# ACTES DES
# 32E JOURNEES D'ETUDES SUR LA PAROLE


ORGANISEES PAR LE LABORATOIRE PAROLE ET LANGAGE
CNRS - AIX-MARSEILLE UNIVERSITE

SOUS L'EGIDE DE L'ASSOCIATION FRANCOPHONE
DE LA COMMUNICATION PARLEE

AIX-EN-PROVENCE 4-8 JUIN 2018

# JEP 2018

# ACTES DES
# 32E JOURNEES D'ETUDES SUR LA PAROLE

4-8 JUIN 2018

AIX-EN-PROVENCE

ORGANISEES PAR LE LABORATOIRE PAROLE ET LANGAGE
CNRS - AIX-MARSEILLE UNIVERSITE

SOUS L'EGIDE DE L'ASSOCIATION FRANCOPHONE
DE LA COMMUNICATION PARLEE

EDITEURS

MARTIN COOKE (ISCA)
BRIGITTE BIGI (LPL)
JOELLE LAVAUD (LPL)

DIFFUSÉ PAR

INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION (ISCA)

HTTPS://WWW.ISCA-SPEECH.ORG/ISCAWEB/INDEX.PHP/ONLINE-ARCHIVE

ASSOCIATION FRANCOPHONE POUR LA COMMUNICATION PARLÉE (AFCP)

HTTP://WWW.AFCP-PAROLE.ORG/

CREDITS IMAGES

CANOPUS I, 1959, VICTOR VASARELY - AVEC L'AUTORISATION DE LA FONDATION VASARELY

SCHEMA DE PRINCIPE DU POLYPHONOMETRE, 1969 - BERNARD TESTON

TRACES IMPRIMES ISSUS DU POLYPHONOMETRE, 1969 - BERNARD TESTON

CONCEPTION COUVERTURE : YOHANN MEYNADIER & CLAUDIA PICHON-STARKE


# Bienvenue aux JEP à Aix-en-Provence

Après les éditions de 1971, 1977, 1986 et 2004 (JEP-TALN à Fès), le Laboratoire Parole et Langage organise, sous l'égide et la caution scientifique de l'Association Francophone de la Communication Parlée (AFCP), la 32è édition des Journées d'Etudes sur la Parole à Aix-en-Provence. Interdisciplinaires, internationales et intergénérationnelles, les JEP offrent aujourd'hui un espace préservé d'échanges scientifiques dans notre langue, de rencontres régulières et conviviales, et d'intégration bienveillante des jeunes chercheurs de notre communauté. Plus qu'une conférence, les JEP sont devenues un rendez-vous attendu de la communauté francophone des recherches scientifiques et technologiques sur la parole depuis maintenant presque 50 ans. On peut même dire depuis 50 ans cette année, si on accepte comme référence « *les journées scientifiques organisées par Max Wajkop à Bruxelles en 1968 pendant les jeux olympiques de Grenoble* », où le projet de voir naître et s'organiser une communauté scientifique francophone de la Parole s'est incarné : « *C'est à l'occasion de cette rencontre bruxelloise, autour de bières, que nous avons pensé à structurer la communauté francophone en développement par des journées d'études annuelles, par un financement spécifique, par une revue européenne, … Il s'agissait de développer une approche pluridisciplinaire coopérative entre ingénieurs-physiciens et phonéticiens de la jeune génération […]. Il s'agissait aussi de trouver une place et de jouer un rôle au sein de la communauté internationale. Les participants aux 'débats imbibés' étaient les suivants (par ordre alphabétique) : Carré, Lancia, Landercy, Paillé, Rossi et Wajskop (Fry, Lane, Mettas étaient plus sérieux). Les journées d'études sont devenues les JEP (en 1970), le financement spécifique s'est appelé Greco (lancé environ 10 ans plus tard avec JP Haton comme directeur) et la revue européenne s'est appelée Speech Communication (éditeur Wajskop)* » (René Carré, 15/05/2018, communication personnelle).

Dès lors, Aix-en-en-Provence a souhaité proposer une édition spéciale des JEP rendant hommage à cette histoire et à ses acteurs appartenant à différentes générations de chercheurs et venant de multiples laboratoires francophones. Le programme comporte ainsi la nouveauté d'une demi-journée spéciale avec deux sessions de conférences centrées sur le rôle joué par les JEP dans l'aventure francophone des recherches sur la parole, revenant sur le développement des instrumentations, des technologies et des recherches en phonétique au cours de presque 50 ans de JEP. A cette occasion, le LPL a aussi voulu célébrer deux figures aixoises emblématiques de l'histoire et l'essor de notre communauté : Bernard Teston et Mario Rossi. Au-delà, nous avons souhaité associer à ce témoignage nos anciens collègues ayant marqué à différentes époques cette histoire. Nous avons donc lancé une vaste invitation à nos 'pères', espérant les voir nombreux réunis lors de cette journée 'historique'. Enfin, à cette occasion, nous avons sollicité les laboratoires pour tenter d'enrichir le site du congrès par une exposition permanente de posters, de documents ou documentaires audiovisuels témoignant du patrimoine historique instrumental, scientifique et humain de la communauté de la Parole.

Une autre spécificité de ces JEP est le renouvellement de l'expérience initiée par Paris en 2016 par l'organisation d'une journée consacrée à des ateliers thématiques satellites aux JEP. Six ateliers sont ainsi proposés, permettant aux participants des JEP (ou non) de prolonger leur programme scientifique en assistant à des rencontres sur des sujets connexes ou plus spécifiques de nos domaines de recherche. Cette année, ces ateliers épousent une initiative d'ouverture auprès de nos collègues didacticiens, deux d'entre eux ayant trait à la parole dans le champ du français langue étrangère.

Cette édition des JEP comporte également quelques autres petites nouveautés. L'une d'elle concerne les sessions orales pour lesquelles le comité de programme a voulu expérimenter d'enrichir chaque session orale par une discussion finale complémentaire, donnant une place élargie aux échanges scientifiques publics. Un grand merci aux présidents de séance qui ont accepté de tenter cette expérience. Par ailleurs, le comité de programme a œuvré pour concevoir des sessions orales thématiques combinant, autant que possible, les travaux de recherche en sciences humaines et en technologies de la parole. En sus, la préoccupation de mieux exposer les travaux des jeunes chercheurs, étudiants ou non statutaires, a également présidé, visant à leur offrir une plus grande visibilité pour soutenir leur intégration dans notre communauté. Deux sessions posters sont, quant à elles, non thématiques afin de privilégier la visibilité du large spectre de nos recherches dans tous les domaines de la parole. Enfin, les Actes des JEP 2018 seront archivés en accès libre non seulement sur le site de l'AFCP, mais également pour la première fois sur celui de l'ISCA.

Dernière 'petite' innovation, les JEP 2018 se voudront économes. Ce sera donc des JEP à 'sacoche recyclable' : pas de sac de congrès, pas d'actes papier, pas de clé USB, pas de mug, pas de blingbling... Que des petites choses, voire des surprises, éphémères et biodégradables, autant que possible !

Mais surtout, les JEP ne seraient plus les JEP si des moments forts en convivialité étaient oubliés. Les 'habitués' le savent bien, l'organisation des JEP répond à un cahier des charges, implicite, mais sévère sur ce point ! La convivialité est le meilleur moyen que les JEP ont su spontanément emprunter pour intégrer, souder, incarner et humaniser notre communauté. L'organisation aixoise des JEP 2018 a tenté de répondre à ce défi, si magistralement relevé par ses prédécesseurs. Ainsi, chaque journée sera conclue par une soirée plus ou moins 'animée', autour d'un apéritif plus ou moins dinatoire ou d'un dîner. La soirée de gala du congrès suivant la journée des sessions 'historiques' sera le point d'orgue festif, musical et pictural de ces JEP.

Pour conclure, nous tenons à remercier ici très chaleureusement les nombreux acteurs et partenaires de cette manifestation : le Laboratoire Parole et Langage, Aix-Marseille Université, le CNRS, la Communauté du Pays d'Aix, l'ILCB/BLRI, l'ISCA, ELRA, Ortolang, la Fondation Vasarely ; et évidemment, l'AFCP, notamment pour l'important travail effectué en tant que comité de programme. En particulier, les organisateurs veulent surtout remercier Jacqueline Vaissière, Didier Demolin, Jean-François Bonastre, Jean-Paul Haton, Christine Meunier et Philippe Martin d'avoir accepté notre sollicitation pour la session 'historique', ainsi que Frédéric Béchet et Noël Nguyen qui orchestreront cette 'cérémonie'. Notre profonde reconnaissance va évidemment à Maxine Eskénazi et Pascal Perrier pour nous honorer d'enrichir nos JEP de leur conférence plénière. Enfin, un sincère remerciement à nos 95 relecteurs mobilisés pour offrir leur expertise aux 183 auteurs de 107 papiers soumis, dont 78 sont présentés dans cette édition (34 oralement et 44 en poster).

A cette heure, venant à près de 120 (et presqu'autant pour la journée des ateliers) d'une trentaine de laboratoires de France, de Belgique, de Suisse, du Canada, voire d'Allemagne et d'Italie, avec x envies d'x idées à échanger, nous vous souhaitons à tous et à toutes la bienvenue à Aix-en-Provence ! Et, surtout, nous espérons vous en voir repartir… en vous disant que vous avez gagné du temps… et surtout du bon temps aux JEP 2018 !

Le comité d'organisation des JEP

le 19 mai 2018, Aix-en-Provence.

# Le mot de la Présidente de l'AFCP

Chers collègues,

C'est sans nul doute l'un des points d'ancrage de la communauté scientifique "Parole" : les JEP, édition 2018, sont sur le point de commencer...

Cette année marque la 32e édition de ces journées d'étude, qui réunissent depuis près de 50 ans les chercheurs en sciences et technologies de la parole pour de fructueux échanges scientifiques et de mémorables moments de convivialité. Après 1971, 1977 et 1986, les JEP reviennent, pour notre plus grand plaisir, dans la magnifique ville d'Aix-en-Provence.

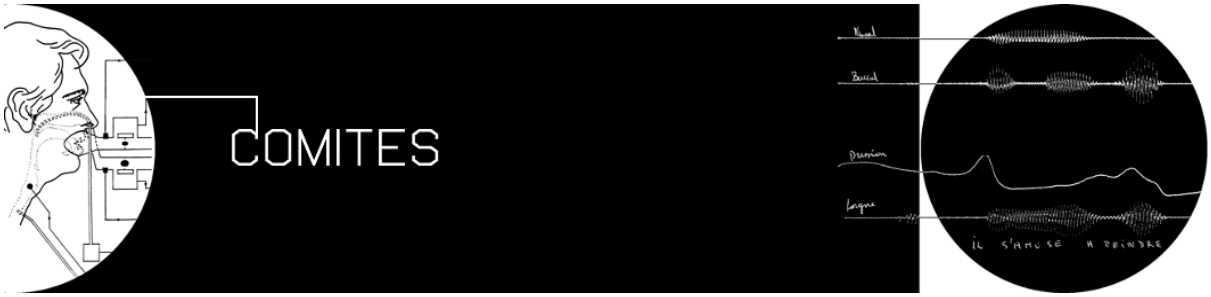
A l'initiative du comité d'organisation local, et avec le soutien de l'AFCP, l'édition 2018 des JEP est placée sous le signe de l'histoire, de la mémoire, et du patrimoine : l'histoire des recherches en parole menées depuis plusieurs décennies dans nos laboratoires, la mémoire des générations de chercheurs qui s'y sont consacrés avec talent et enthousiasme, le patrimoine scientifique et instrumental qu'ils nous ont légués. Ce coloriage thématique se déclinera à travers divers événements qui jalonneront toute la semaine des JEP (sessions orales, exposition et installations, ateliers, etc.), avec en point d'orgue la journée du 6 juin, où nous rejoindrons divers invités de marque, et au cours de laquelle un hommage particulier sera rendu à deux figures aixoises qui ont particulièrement travaillé à l'essor de la communauté "Parole" : Bernard Teston et Mario Rossi.

Cet ancrage des JEP 2018 dans ce qui constitue le fondement de notre communauté scientifique est d'autant plus opportun que, sous bien des aspects, nous sommes à la croisée des chemins. Face à la pression croissante en faveur de la publication référencée en anglais, de la valorisation immédiate des résultats, et dans un contexte général de compétitivité accrue, d'hyperspécialisation thématique, et de courses aux indicateurs de visibilité, les JEP peuvent apparaître comme un îlot de résistance. Celui d'une conférence à taille humaine, où les chercheurs francophones prennent le temps d'échanger dans leur langue maternelle à propos de travaux de qualité, consacrés par des articles complets. Cette année encore, le comité de programme a établi le programme scientifique de façon à décloisonner, à favoriser les échanges et les regards croisés : absence de sessions orales parallèles, sessions poster non thématiques, sessions orales regroupant autour de thématiques communes des travaux issus des sciences humaines et sociales ainsi que des sciences et technologies de l'information, etc. Au sein du CA de l'AFCP, nous croyons en l'esprit JEP, et nous consacrons nos énergies et nos ressources à assurer un avenir à cet événement-phare pour notre communauté !

Je terminerai ce billet en remerciant tous ceux – et ils sont nombreux – qui ont rendu cet événement possible. Merci aux évaluateurs (près d'une centaine), leurs relectures attentives ont permis d'assurer la qualité des communications finalement retenues. Merci aux membres du comité de programme, leur investissement dans le long processus d'évaluation, de sélection et de coordination a permis d'aboutir à un programme scientifique équilibré et attrayant. Un tout grand merci, enfin, aux membres du comité d'organisation local du Laboratoire Parole et Langage d'Aix, en particulier Alain Ghio, Christine Meunier et Yohann Meynadier, qui ont œuvré depuis plusieurs mois à tous les niveaux afin de faire de cette organisation une réussite. Si ces 32e JEP sont d'ores et déjà un succès, c'est avant tout grâce à eux !

Je vous souhaite donc à tous d'excellentes JEP2018, riches de découvertes et d'échanges, scientifiques et personnels.

Véronique Delvaux
Présidente du Comité de Programme
Présidente du Conseil d'Administration de l'AFCP




# Comité local d'organisation

Laboratoire Parole et Langage, Aix-en-Provence

Carine André
Brigitte Bigi
Sébastien Bermond
Philippe Blache
Christian Cavé
Cyril Deniaud
**Alain Ghio, co-présidence**
Aurélie Goujon
Isabelle Guaïtella
Sophie Herment
Muriel Lalain
Rémi Lamarque
Joëlle Lavaud
Amélie Leconte
Frédéric Lefèvre
Thierry Legou
**Stéphanie Desous, secrétariat général**
Anna Marczyk
Amandine Michelas
Christine Meunier
**Yohann Meynadier, co-présidence**
Nadia Monségu
Claudia Pichon-Starke
Gilles Pouchoulin
Marion Tellier
Anne Tortel

# Comité de programme

conseil d'administration de l'AFCP

**Martine Adda-Decker, LPP, CNRS, U. Paris 3, présidente adjointe**
Jean-François Bonastre, LIA, U. Avignon
Fethi Bougares, LIUM, U. du Maine
Philippe Boula de Mareüil, LIMSI, CNRS, Paris
Hervé Bredin, LIMSI, CNRS, Paris
Olivier Crouzet, LLing, CNRS, U. Nantes
Elisabeth Delais-Roussarie, LLing, CNRS, U. Nantes
**Véronique Delvaux, IRSTL, U. Mons, présidente**
Camille Fauth, LiLPa, U. Strasbourg
Emmanuel Ferragne, CLILLAC-ARP, U. Paris 7
Cécile Fougeron, LPP, CNRS, U. Paris 3
Corinne Fredouille, LIA, U. Avignon
Alain Ghio, LPL, CNRS, U. Aix-Marseille
Camille Guinaudeau, LIMSI, CNRS, Paris
Anne Guyot-Talbot, CLILLAC-ARP, U. Paris 7
Bernard Harmegnies, IRSTL, U. Mons
Nathalie Henrich Bernardoni, Gipsa-Lab, CNRS, U. Grenoble Alpes
Bassam Jabaian, LIA, U. Avignon
David Langlois, LORIA, CNRS, U. Lorraine
Yves Laprie, LORIA, CNRS, U. Lorraine
Anthony Larcher, LIUM, U. du Maine
Gwénolé Lecorvé, IRISA, U. Rennes
Benjamin Lecouteux, LIG, CNRS, U. Grenoble Alpes
Georges Linares, LIA, U. Avignon
Damien Lolive, IRISA, U. Rennes
Julie Mauclair, IRIT, CNRS, U. Toulouse
**Christine Meunier, LPL, CNRS, U. Aix-Marseille, invitée**
Yohann Meynadier, LPL, CNRS, U. Aix-Marseille
Slim Ouni, LORIA, CNRS, U. Lorraine
Thomas Pellegrini, IRIT, CNRS, U. Toulouse
François Portet, LIG, CNRS, U. Grenoble Alpes
Fabiàn Santiago, SFL, CNRS, U. Paris 8
Christophe Savariaux, Gipsa-Lab, CNRS, U. Grenoble Alpes
Nathalie Vallée, Gipsa-Lab, CNRS, U. Grenoble Alpes
Ioana Vasilescu, LIMSI, CNRS, Paris

# Comité scientifique

Gilles Adda, LIMSI, CNRS, Paris
Martine Adda-Decker, LPP, CNRS, U. Paris 3
Angélique Amelot, LPP, CNRS, U. Paris 3
Nicolas Audibert, LPP, CNRS, U. Paris 3
Pierre Badin, Gipsa-Lab, CNRS, U. Grenoble Alpes
Melissa Barkat, ISEM, CNRS, U. Montpellier
Claude Barras, LIMSI, CNRS, Paris
Jean-François Bonastre, LIA, U. Avignon
Anne Bonneau, LORIA, CNRS, U. Lorraine
Fethi Bougares, LIUM, U. du Maine
Philippe Boula de Mareüil, LIMSI, CNRS, Paris
Hervé Bredin, LIMSI, CNRS, Paris
Lise Crevier-Buchman, LPP, CNRS, U. Paris 3
Olivier Crouzet, LLing, CNRS, U. Nantes
Elisabeth Delais-Roussarie, LLing, CNRS, U. Nantes
Véronique Delvaux, IRSTL, U. Mons
Didier Demolin, LPP, CNRS, U. Paris 3
Ivana Didirkovà, Praxiling, CNRS, U. Montpellier
Mariapaola D'Imperio, LPL, CNRS, U. Aix-Marseille
Christelle Dodane, Praxiling, CNRS, U. Montpellier
Sophie Dufour, LPL, CNRS, U. Aix-Marseille
Frédéric Elisei, Gipsa-Lab, CNRS, U. Grenoble Alpes
Simone Falk, LPP, CNRS, U. Paris 3
Camille Fauth, LiLPa, U. Strasbourg
Gang Feng, Gipsa-Lab, CNRS, U. Grenoble Alpes
Emmanuel Ferragne, CLILLAC-ARP, U. Paris 7
Isabelle Ferrané, IRIT, CNRS, U. Toulouse
Cécile Fougeron, LPP, CNRS, U. Paris 3
Corinne Fredouille, LIA, U. Avignon
Cédric Gendrot, LPP, CNRS, U. Paris 3
Silvain Gerber, Gipsa-Lab, CNRS, U. Grenoble Alpes
Sahar Ghannay, LIUM, U. du Maine
Alain Ghio, LPL, CNRS, U. Aix-Marseille
Camille Guinaudeau, LIMSI, CNRS, Paris
Patrice Guyot, IRIT, CNRS, U. Toulouse
Anne Guyot-Talbot, CLILLAC-ARP, U. Paris 7
Pierre Hallé, LPP, CNRS, U. Paris 3
Bernard Harmegnies, IRSTL, U. Mons
Nathalie Henrich, Gipsa-Lab, CNRS, U. Grenoble Alpes
Sophie Herment, LPL, CNRS, U. Aix-Marseille
Fabrice Hirsch, Praxiling, CNRS, U. Montpellier
Kathy Huet, IRSTL, U. Mons
Bassam Jabaian, LIA, U. Avignon
Thomas Jauriberry, LiLPa, U. Strasbourg
Takeki Kamiyama, LPP, CNRS, U. Paris 3
Barbara Kühnert, LPP, CNRS, U. Paris 3
Imed Laaridh, IRIT, CNRS, U. Toulouse
Muriel Lalain, LPL, CNRS, U. Aix-Marseille
David Langlois, LORIA, CNRS, U. Lorraine

Yves Laprie, LORIA, CNRS, U. Lorraine
Anthony Larcher, LIUM, U. du Maine
Gwénolé Lecorvé, IRISA, U. Rennes
Benjamin Lecouteux, LIG, CNRS, U. Grenoble Alpes
Thierry Legou, LPL, CNRS, U. Aix-Marseille
Georges Linares, LIA, U. Avignon
Damien Lolive, IRISA, U. Rennes
Paolo Mairano, LFSAG, Université de Turin, Italie
Anna Marczyk, LPL, CNRS, U. Aix-Marseille
Julie Mauclair, IRIT, CNRS, U. Toulouse
Christine Meunier, LPL, CNRS, U. Aix-Marseille
Julien Meyer, Gipsa-Lab, CNRS, U. Grenoble Alpes
Yohann Meynadier, LPL, CNRS, U. Aix-Marseille
Alexis Michaud, LACITO, CNRS, U. Paris 3
Amandine Michelas, LPL, CNRS, U. Aix-Marseille
Noël Nguyen, LPL, CNRS, U. Aix-Marseille
Slim Ouni, LORIA, CNRS, U. Lorraine
Thomas Pellegrini, IRIT, CNRS, U. Toulouse
Pascal Perrier, Gipsa-Lab, CNRS, U. Grenoble Alpes
Myriam Piccaluga, IRSTL, U. Mons
Claire Pillot-loiseau, LPP, CNRS, U. Paris 3
François Portet, LIG, CNRS, U. Grenoble Alpes
Rachid Ridouane, LPP, CNRS, U. Paris 3
Albert Rilliard, LIMSI, CNRS, Paris
Solange Rossato, LIG, CNRS, U. Grenoble Alpes
Halima Sahraoui, Octogone-Lordat, U. Toulouse
Fabiàn Santiago, SFL, CNRS, U. Paris 8
Marc Sato, LPL, CNRS, U. Aix-Marseille
Christophe Savariaux, Gipsa-Lab, CNRS, U. Grenoble Alpes
Jean Schoentgen, Ecole Polytechnique, U. Libre Bruxelles
Christine Sénac, IRIT, CNRS, U. Toulouse
Rudolph Sock, LiLPa, U. Strasbourg
Anne Tortel, LPL, CNRS, U. Aix-Marseille
Gabor Turcsan, LPL, CNRS, U. Aix-Marseille
Jacqueline Vaissière, LPP, CNRS, U. Paris 3
Nathalie Vallée, Gipsa-Lab, CNRS, U. Grenoble Alpes
Ioana Vasilescu, LIMSI, CNRS, Paris
Béatrice Vaxelaire, LiLPa, U. Strasbourg
Clemence Verhaegen, IRSTL, U. Mons
Coriandre Vilain, Gipsa-Lab, CNRS, U. Grenoble Alpes
Pauline Welby, LPL, CNRS, U. Aix-Marseille
Jane Wottawa, LIMSI, CNRS, Paris
Naomi Yamaguchi, LPP, CNRS, U. Paris 3
Hiyon Yoo, LLF, CNRS, U. Paris 7


# Remerciements

Nous souhaitons remercier les organisateurs pour leur soutien logistique et financier :

UMR 7309 • CNRS • AMU
PAROLE ET LANGAGE

AFCP

Ainsi que nos partenaires pour leur soutien et leur confiance :

Aix*Marseille université
Initiative d'excellence

cnrs
dépasser les frontières

EUROPEAN LANGUAGE RESOURCES ASSOCIATION
ELRA

Ortolang
Open Resources and Tools for LANGuage

ILCB
Institute of Language Communication and the Brain

METROPOLE AIX MARSEILLE PROVENCE

FONDATION VASARELY


# Lundi 4 juin

| | | |
|---|---|---|
| **12:00-13:30** | Inscriptions (Hall d'accueil) | |
| **13:30-14:00** | Ouverture des JEP 2018 (Amphi) | |
| **14:00-15:20** | VOIX (Amphi) - M. Garnier ; T. Pellegrini | |
| | **Conversion d'Identité de la Voix Chantée par Sélection et Concaténation d'Unités Spectrales**<br>*Obin Nicolas, Pascal Pham, Roebel Axel* | 1 |
| | **Mesure de similarité fondée sur des réseaux de neurones siamois pour le doublage de voix**<br>*Gresse Adrien, Dufour Richard, Labatut Vincent, Rouvier Mickael, Bonastre Jean-François* | 10 |
| | **Doubler les consonnes en chant baroque français : un cas de gémination expressive ?**<br>*Pillot-Loiseau Claire, Schweitzer Claudia, Dodane Christelle, Romeo Alice, Turco Giuseppina* | 19 |
| **15:20-16:40** | LOCUTEURS (Amphi) - V. Delvaux ; I. Ferrané | |
| | **Comparaison des voix dans le cadre judiciaire : influence du contenu phonétique**<br>*Ajili Moez, Bonastre Jean-François, Ben Kheder Waad, Rossato Solange, Kahn Juliette* | 28 |
| | **Suivre le rythme de tes paroles**<br>*Rossato Solange, Zhang Da, Ajili Moez, Bonastre Jean-François* | 37 |
| | **Variabilité inter et intra locuteurs de mesures spectrales et prosodiques en parole lue**<br>*Gendrot Cédric, Chignoli Gabriele, Audibert Nicolas, Fougeron Cécile* | 46 |
| **16:40-17:10** | Pause café (Patio) | |
| **17:10-18:55** | VARIATION (Amphi) - E. Delais ; C. Gendrot | |
| | **Le /R/ « roulé » en français et dans quelques langues régionales de France**<br>*Premat Timothée, Boula De Mareüil Philippe* | 55 |
| | **Quand les voyelles longues et brèves ne tiennent pas en place : la qualité vocalique en allemand L2**<br>*Wottawa Jane, Adda-Decker Martine* | 64 |
| | **Étude acoustique de voyelles tenues produites par des patients** | 72 |

| | | |
|---|---|---|
| | **glossectomisés suite à un cancer endo-bucal**<br>*Zaouali Hasna, Vaxelaire Béatrice, Debry Christian, Bronner Guy, Sock Rudolph* | |
| | **Efforts de production de parole chez les personnes qui bégaient**<br>*Garnier Maëva, Da Fonseca Anaïs, Savariaux Christophe, Cattelain Thibault* | 80 |
| **18:55-19:10** | Point infos | |
| **20:00-21:30** | Pot d'accueil au Pavillon Vendôme | |

# Mardi 5 juin

| | | |
|---|---|---|
| **09:00-10:00** | Conférence plénière (Amphi) - Maxine ESKENAZI - (M. Adda-Decker)<br>**Les systèmes de dialogue oral : avancées et limites** | |
| **10:00-10:30** | Pause café (Patio) | |
| **10:30-12:15** | CORRECTIONS ET INCERTITUDES (Amphi) - F. Pellegrino ; B. Favre | |
| | **Segmentation et Regroupement en Locuteurs : comment évaluer les corrections humaines**<br>*Broux Pierre-Alexandre, Doukhan David, Petitrenaud Simon, Meignier Sylvain, Carrive Jean* | 89 |
| | **Transcription phonétique automatique pour la synthèse de la parole**<br>*Vythelingum Kévin, Estève Yannick, Rosec Olivier* | 98 |
| | **Analyse électromyographique de la production des plosives labiales : Enjeux méthodologiques**<br>*Cattelain Thibault, Garnier Maëva, Savariaux Christophe, Gerber Silvain, Perrier Pascal* | 107 |
| | **L'incidence de la correction phonétique sur l'acquisition des voyelles en langue étrangère : étude de cas d'anglophones apprenant le français**<br>*Alazard-Guiu Charlotte, Santiago Fabiàn, Mairano Paolo* | 116 |
| **12:15-13:45** | Déjeuner (Restaurant Universitaire les Fenouillères) | |
| **14:00-15:30** | POSTER 1 (Patio) - A. Goujon ; R. Lamarque ; A. Leconte ; G. Pouchoulin | |
| | **Ambiguïté temporaire des obstruantes voisées en parole chuchotée**<br>*Meynadier Yohann, Dufour Sophie* | 125 |
| | **Codage efficace à débit variable basé sur la quantification vectorielle à divisions commutées : Application aux paramètres ISF en large bande**<br>*Cheraitia Salah-Eddine, Bouzid Merouane, Meziane Nacéra* | 134 |
| | **Développement de la parole et de la mastication : Evolution de la durée des cycles oscillatoires mandibulaires observés entre 8 et 14 mois chez 4 enfants québécois**<br>*Lemarchand Leslie, MacLeod A.N. Andrea, Canault Mélanie, Kern Sophie* | 142 |
| | **Effet de la position de la syllabe sur la réalisation acoustique des consonnes finales du thaï**<br>*Yamlamai Nicha, Tran Thi Thuy Hien* | 151 |

| | | |
|---|---|---|
| | **Variabilité du geste linguo-palatal. Le cas du russe**<br>*Biteeva Lecocq Ekaterina, Vallée Nathalie, Gerber Silvain, Savariaux Christophe* | 294 |
| | **Vers un modèle du « toucher vocal » pour la communication ubiquïte**<br>*Davat Ambre, Aubergé Véronique, Feng Gang* | 303 |
| | **L'opposition de voisement chez les apprenants syriens de FLE**<br>*Abou Haidar Laura* | 312 |
| **15:30-16:00** | Pause café (Patio) | |
| **16:00-17:45** | INTERACTION (Amphi) - R. Bertrand ; L. Besacier | |
| | **Gestes et prosodie dans la parole aphasique non fluente**<br>*Ferré Gaëlle* | 320 |
| | **« Tout ça c'est abstrait » : Comment le degré d'abstraction d'un mot expliqué affecte-t-il la parole multimodale ?**<br>*Tellier Marion, Stam Gale, Ghio Alain* | 329 |
| | **Effet de la situation de parole sur la variabilité des voyelles en français**<br>*Lancien Mélanie, Audibert Nicolas, Fougeron Cécile* | 338 |
| | **Évaluation de l'adaptation par renforcement d'un générateur en langage naturel neuronal pour le dialogue homme-machine**<br>*Riou Matthieu, Jabaian Bassam, Huet Stéphane, Lefèvre Fabrice* | 347 |
| **17:45-18:00** | Point infos | |
| **18:00-19:00** | Assemblée Générale de l'AFCP (Amphi) - V. Delvaux ; M. Adda-Decker | |
| **19:30-22:00** | Evénement social - Apéritif Boules à l'anis aux 2 Frères | |

# Mercredi 6 juin

| | | |
|---|---|---|
| **09:00-10:00** | PAROLE A L'HISTOIRE (Amphi) - J. Vaissière ; S. Lienard | |
| | **Joseph Fourier : quelques points de repère, dans un héritage exceptionnel**<br>(communication invitée pour les commémorations nationales '2018, année Fourier')<br>*Bimbot Frédérique* | |
| | **L'histoire des alphabets phonétiques du XVIIIe jusqu'à l'API**<br>*Schweitzer Claudia, Dodane Christelle, Lazar Jan* | 356 |
| | **Une histoire des JEP : 50 ans d'études sur la parole**<br>*Delvaux Véronique, Luxardo Giancarlo, Hirsch Fabrice* | 365 |
| **10:00-10:30** | Pause café (Patio) | |
| **10:30-12:15** | SESSION SPECIALE 'HISTORIQUE' (Amphi) - N. Nguyen ; F. Béchet<br>en hommage à Bernard Teston et Mario Rossi | |
| | **Du kymographe à Eva : petite histoire de l'instrumentation phonétique en France. (Hommage à Bernard Teston)**<br>*D. Demolin, J. Vaissière* | |
| | **50 ans de communication parlée en France : aspects technologiques** | |

| | | |
|---|---|---|
| | *JF Bonastre, JP Haton* | |
| | **La phonétique segmentale et prosodique en 50 ans de JEP**<br>*C. Meunier, P. Martin* | |
| | **Hommage à Mario Rossi**<br>*A. Ghio* | |
| **12:15-13:45** | Déjeuner (Restaurant Universitaire les Fenouillères) | |
| **14:00-15:45** | PERCEPTION (Amphi) - B. Harmegnies ; C. Fredouille | |
| | **Peut-on distinguer perceptivement huit accents régionaux en français parlé en Europe ? Une réponse à base de *crowdsourcing***<br>*Avanzi Mathieu, Boula De Mareüil Philippe* | 374 |
| | **L'information accentuelle est-elle représentée dans le lexique mental des locuteurs du français ?**<br>*Michelas Amandine, Dufour Sophie* | 383 |
| | **Évaluations perceptive et automatique de l'intelligibilité de la parole dégradée par simulation de la surdité professionnelle**<br>*Laaridh Imed, Tardieu Julien, Magnen Cynthia, Gaillard Pascal, Farinas Jérôme, Pinquier Julien* | 392 |
| | **Perception des consonnes et voyelles nasales en parole vocodée : Analyse de la contribution des niveaux de résolution spectrale et temporelle**<br>*Crouzet Olivier* | 401 |
| **15:45-16:15** | Pause café (Patio) | |
| **16:15-17:35** | SIGNAUX SOCIAUX (Amphi) - J. Revis ; Y. Estève | |
| | **Réduction de la coarticulation et vieillissement**<br>*D'Alessandro Daria, Fougeron Cécile* | 410 |
| | **Voix et sélection sexuelle : une approche interdisciplinaire**<br>*Suire Alexandre, Raymond Michel, Barkat-Defradas Melissa* | 419 |
| | **L'abaissement de la fréquence fondamentale comme pratique de séduction**<br>*Arnold Aron* | 424 |
| **17:35-18:00** | Point infos | |
| **20:30-02:00** | Soirée de Gala à la Fondation Vasarely | |

# Jeudi 7 juin

| | | |
|---|---|---|
| **10:00-11:00** | Conférence plénière (Amphi) - Pascal PERRIER - (J. Schoentgen)<br>**Buts moteurs de la production de la parole : apport de la modélisation biomécanique et des expériences de perturbations** | |
| **11:00-12:20** | ARTICULATION (Amphi) - P. Perrier ; F. Hirsch | |
| | **Evolution des habiletés articulatoires au stade du babillage : le timing des syllabes CV**<br>*Canault Mélanie, Kern Sophie, Yamaguchi Naomi, Dos Santos Christophe,* | 433 |

# Vendredi 8 juin

| | | |
|---|---|---|
| **09:00-09:30** | Ouverture des Ateliers satellites | |
| **09:30-11:00** | Atelier 1 - **PinPex** - Y. Meynadier, F. Hirsch | |
| **09:30-11:00** | Atelier 2 - **Interactions sociales conversationnelles** - T. Chaminade, F. Lefèvre, N. Nguyen, M. Ochs | |
| **09:30-11:00** | Atelier 3 - **Transposer les études en linguistique de l'oral** - E. Oursel, C. Etienne, E. Jouin-Chardon, V. André, C. David | |
| **11:00-11:30** | Pause café (Patio) | |
| **11:30-13:00** | Atelier 1 - **PinPex** | |
| **11:30-13:00** | Atelier 2 - **Interactions sociales conversationnelles** | |
| **11:30-13:00** | Atelier 3 - **Transposer les études en linguistique de l'oral** | |
| **13:00-14:00** | Pause libre | |
| **14:00-15:30** | Atelier 4 - **Regards croisés sur l'apraxie** - A. Marczyk, V. Sabadell, C. Meunier, A. Fasola, C. Verhaegen, L. Baqué, A. Rosas, T. Prince | |
| **14:00-15:30** | Atelier 5 - **Groupe d'instrumentation de la parole** - Legou, T. Cattelain, F. Silva | |
| **14:00-15:30** | Atelier 6 - **La parole dans l'espace de la classe** - V. Bourhis, R. Gagnon | |
| **15:30-16:00** | Pause café (Patio) | |
| **16:00-17:30** | Atelier 4 - **Regards croisés sur l'apraxie** | |
| **16:00-17:30** | Atelier 5 - **Groupe d'instrumentation de la parole** | |
| **16:00-17:30** | Atelier 6 - **La parole dans l'espace de la classe** | |
| **17:30-18:00** | Clôture des Ateliers satellites | |


**Les systèmes de dialogue oral : avancées et limites**

Maxine ESKENAZI
*Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA*

Intelligent agents have entered the marketplace and many people use them on a daily basis. We will discuss what works in these systems such as the far field microphone developments and question-answer pairs. We will also try to shed some light on why this has attracted so many users. Then we will look under the hood at the classic spoken dialog system pipeline architecture and the basics of how it works. Following this, we will examine the architecture of a multi-domain portal. Since our multi-domain portal aggregates many heterogeneous systems, we will discuss the issues involved in its creation and maintenance. Due to significant public awareness of these systems, intelligent agents are at a crucial point in their development. Our systems can continue to be accepted and become ubiquitous in our lives. But some issues could quickly destroy public confidence and consequently have an effect on support for research in related areas. There are major problems that have come to light as our research advances and as agents are used by so many people. One example is the recent privacy breaches. We will discuss the main issues that intelligent agents face today and the very promising research paths that await us.

**Buts moteurs de la production de la parole :**
**apport de la modélisation biomécanique et des expériences de perturbations**

Pascal PERRIER
*Gipsa Lab, Ecole Polytechnique & CNRS, Grenoble, France*

La question de la nature des buts moteurs en parole est ancienne, et remonte sans doute à l'origine des travaux en phonétique expérimentale. Elle est assurément au coeur des vifs débats que notre domaine a connus dans la fin des années 80 et au cours des années 90 sur la nature des propriétés physiques qui pour les uns constituent l'invariant porteur de l'information linguistique (voir Perkell & Klatt, 1986) et pour les autres sont les éléments qui caractérisent la monnaie d'échange (Goldstein & Fowler, 2012) entre un locuteur et son ou ses interlocuteur(s). Les enjeux principaux étaient alors de savoir si les buts moteurs de la parole étaient plutôt auditifs, plutôt articulatoires ou plutôt moteurs. Aujourd'hui, il est généralement admis que c'est probablement un peu de tout cela, et l'hypothèse du caractère multimodal des buts moteurs est le plus souvent bien acceptée (voir par exemple les propositions à la base du modèle de production aujourd'hui le plus célèbre, DIVA, élaboré par Guenther et son équipe, Guenther & Vladusich, 2012). Mais des questions restent posées qui concernent par exemple la possibilité d'une hiérarchie entre les différentes modalités et la nature temporelle de ces buts. Les progrès techniques réalisés d'une part dans les outils de modélisation numérique, qui ont permis d'avancer de manière très significatives dans le développement de modèles physiques réalistes de la production de la parole, et d'autre part dans les outils temps réel de traitement du signal, qui ont ouvert le champ à de nouveaux types de perturbation en ligne de la production de la parole, ont permis l'émergence de nouvelles méthodologies pour tester plus précisément les hypothèses théoriques dans ce domaine.
C'est à la présentation de certains de ces travaux, issus pour partie de notre équipe à GIPSA-lab, mais aussi tirés de la littérature, et à leur analyse que sera consacré cet exposé. Ils reposent pour partie sur des simulations obtenues avec des modèles biomécaniques des articulateurs de la parole et pour partie sur des expériences d'apprentissage moteur de la production de la parole impliquant des perturbations de l'articulation de la parole et/ou du retour auditif.

Depuis le milieu des années 90, nous travaillons à l'Institut de la Communication Parlée, puis à GIPSA-lab, sur le développement de modèles biomécaniques réalistes des articulateurs orofaciaux, que nous exploitons pour connaître les contraintes que la biomécanique impose à la production de la parole et déterminer leur rôle dans la mise en forme temporelle des signaux articulatoires et acoustiques de la parole. Nous avons pu ainsi montrer que du fait des propriétés physiques de l'appareil de production de la parole (1) des trajectoires articulatoires complexes peuvent prendre forme à partir de buts moteurs discrets sans nécessiter des stratégies de planification gestuelle ou de contrôle moteur sophistiquées (Perrier et al., 2003) ; (2) des propriétés cinématiques fortes émergent naturellement (Perrier et Fuchs, 2008). Nous, ainsi que des collègues dans d'autres laboratoires exploitant des modèles similaires, ont aussi pu montrer que les stratégies articulatoires spécifiques, caractéristiques de certains sons de parole, pourraient émerger d'objectifs dans le domaine auditif, du fait de l'existence de propriétés biomécaniques spécifiques facilitant la réalisation de ces objectifs auditifs (Nazari et al., 2013 ; Stavness et al. 2012).

Nous présenterons tout d'abord les résultats de certaines des expériences que nous avons réalisées dans le laboratoire et dans le cadre de collaborations internationales (Savariaux et al., 1995, Ménard et al., 2008, Brunner et al., 2011) qui nous ont conduits à faire l'hypothèse que la modalité auditive prime dans la spécification des buts de la parole, avec cependant une influence des contraintes somatosensorielles qui ne peuvent être ignorées. Puis nous présenterons les travaux expérimentaux récents, réalisés dans d'autres laboratoires, et en particulier dans celui de David Ostry à Montréal (Lametti et al., 2012, 2014), qui mettent en jeu des perturbations en ligne du retour auditif au cours de la production de la parole. Ces perturbations combinées à des perturbations de l'articulation ont largement questionné notre hypothèse sur la hiérarchie de la modalité auditive. Nous les expliquerons plus en détail.

Nous discuterons l'ensemble de ces résultats, parfois en exploitant un modèle bayésien du contrôle moteur de la parole récemment développé à GIPSA-lab (Patri et al., 2015, 2018), essentiellement par rapport aux deux hypothèses qui auront été reprises tout au long de cet exposé : la nature temporelle discrète des buts moteurs, et la hiérarchie de la modalité auditive dans la spécification de ces buts.

BRUNNER J., HOOLE P., PERRIER P. (2011). Adaptation strategies in perturbed /s/. Clinical Linguistics & Phonetics, 25(8), 705-724.

GOLDSTEIN L., FOWLER C.A (2003). Articulatory phonology: A phonology for public language use. In N.O. Schiller & A.S. Meyer (Eds.) Phonetics and phonology in language comprehension and production: Differences and similarities (pp. 159-207). Berlin, Allemagne : Walter de Gruyter

GUENTHER F.H., VLADUSICH T. (2012). A neural theory of speech acquisition and production. Journal of Neurolinguistics, 25(5), 408-422.

LAMETTI D. R., NASIR S. M., OSTRY D. J. (2012). Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. Journal of Neuroscience, 32(27), 9351-9358.

LAMETTI D. R., ROCHET-CAPELLAN A., NEUFELD E., SHILLER D. M., OSTRY D. J. (2014). Plasticity in the human speech motor system drives changes in speech perception. Journal of Neuroscience, 34(31), 10339-10346.

MENARD, L., PERRIER, P., AUBIN, J., SAVARIAUX, C., & THIBEAULT, M. (2008). Compensation strategies for a lip-tube perturbation of French [u]: An acoustic and perceptual study of 4-year-old children. Journal of the Acoustical Society of America, 124(2), 1192-1206.

NAZARI M.A., PERRIER P., CHABANAS M., PAYAN, Y. (2011). Shaping by stiffening: a modeling study for lips. Motor control, 15(1), 141-168.

PATRI J. F., DIARD J., PERRIER P. (2015). Optimal speech motor control and token-to-token variability: a Bayesian modeling approach. Biological cybernetics, 109(6), 611-626.

PATRI J. F., PERRIER P., SCHWARTZ J. L., DIARD, J. (2018). What drives the perceptual change resulting from speech motor adaptation? Evaluation of hypotheses in a Bayesian modeling framework. PLoS computational biology, 14(1), e1005942.

PERKELL J.S., KLATT, D.H. (1984) (Eds). Invariance and Variability in Speech Processes. Lawrence Erlbaum Associates

PERRIER P., PAYAN Y., ZANDIPOUR M., PERKELL, J. (2003) Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study. Journal of the Acoustical Society of America, 114(3), 1582-1599

PERRIER P., FUCHS, S. (2008) Speed–Curvature Relations in Speech Production Challenge the ⅓ Power Law. J Neurophysiology, 100, 1171-1183.

SAVARIAUX C., PERRIER P., ORLIAGUET J.P. (1995) Compensation Strategies for a Lip-tube Perturbation of the Rounded Vowel [u]. Journal of the Acoustical Society of America, 98 (5), 2428–2442.

STAVNESS I., GICK B., DERRICK D., FELS S. (2012). Biomechanical modeling of English/r/variants. The Journal of the Acoustical Society of America, 131(5), EL355-EL360.

**Joseph Fourier : quelques points de repère, dans un héritage exceptionnel**

Frédéric BIMBOT
*IRISA/CNRS*

En mars 1768 naissait Joseph Fourier, mathématicien et physicien français qui allait découvrir une méthode de décomposition de fonctions en séries trigonométriques dont l'utilité, 250 ans plus tard, n'est plus à prouver ! A l'occasion de cet anniversaire inscrit aux commémorations nationales 2018, cet exposé présentera brièvement quelques points de repères saillants dans la vie et l'oeuvre scientifique de ce savant dont l'héritage exceptionnel est omniprésent, dans nos disciplines et dans beaucoup d'autres.


ARTICLES
Nasal
Buccal
Pression
Langue
IL S'AMUSE À PEINDRE

XXXIIe Journées d'Études sur la Parole
4-8 juin 2018, Aix-en-Provence, France

# Conversion d'Identité de la Voix Chantée par Sélection et Concaténation d'Unités Spectrales

Nicolas Obin, Pascal Pham, Axel Roebel
IRCAM, Sorbonne Université, CNRS, Paris, France
{nobin, pham, roebel}@ircam.fr

## RÉSUMÉ

Cet article présente un algorithme de sélection d'unités spectrales pour la conversion de l'identité de la voix chantée à partir de bases de données non parallèles. Les algorithmes de conversion basés sur des unités de parole présentent des avantages importants pour la conversion de l'identité vocale : la conversion vocale par sélection d'unités permet la préservation des caractéristiques originales de la voix cible, en utilisant des unités réelles ; et la segmentation en unités linguistiques permet d'apprendre la conversion à partir d'enregistrements de la voix cible non nécessairement alignés avec ceux de la voix source. La contribution principale de cet article est de réaliser la sélection des unités spectrales de la voix cible en fonction de plusieurs facteurs : acoustique, linguistique (phonèmes) et musicaux (hauteur, intensité et durée). Pour ce faire, la sélection de la séquence d'unités d'enveloppe spectrale est établie comme un problème d'optimisation à partir d'une fonction de coût multiple qui comprend la distorsion spectrale des chanteurs source et cible ainsi que les différences de hauteur, d'intensité et de durée des unités spectrales correspondantes. L'objectif est de guider la sélection vers des enveloppes spectrales du chanteur cible partageant un contexte musical similaire avec celles du chanteur source. Il est montré lors d'une expérience perceptive que l'algorithme proposé améliore le naturel de la conversion et la similarité avec la voix cible.

## ABSTRACT

This paper presents a unit-selection algorithm for non-parallel singing voice conversion. Unit-based algorithms presents important advantages for voice conversion : the speech segmentation into linguistic units allows the possibility to learn the conversion from on-the-fly databases of the target voice not necessarily aligned to the source voice, and unit-selection voice conversion allows the preservation of the original characteristics of the target voice, by using real units. The main idea of this paper is that the spectral envelopes of a speaker vary according to multiple factors : linguistics (phonemes), and musical (pitch, intensity, and duration). Accordingly, the selection of the sequence of spectral envelope units is established as a multi-target optimization problem, including the spectral distortion of the source and target singers, and the pitch, intensity, and duration differences of the corresponding spectral envelopes. The objective is to guide the selection towards spectral envelopes of the source and target singers sharing a similar musical context. It is shown that the proposed algorithm improves conversion naturalness and target similarity.

MOTS-CLÉS : conversion de l'identité vocale, voix chantée, conversion non-parallèle, sélection d'unités, optimisation multi-cible.

KEYWORDS: voice conversion, singing voice, non-parallel conversion, unit-selection, multi-target


10.21437/JEP.2018-1

optimization.

# 1 Introduction

La conversion d'identité vocale consiste à modifier la voix d'un locuteur source afin d'être perçue comme celle d'un locuteur cible. Grâce aux avancées récentes, la conversion vocale a considérablement gagné en popularité et en qualité au cours des dernières années, menant notamment aux premières compétitions internationales sur la conversion d'identité vocale (Toda *et al.*, 2016; Lorenzo-Trueba *et al.*, 2018), et avec son extension à la conversion vocale entre des langues différentes (Sündermann *et al.*, 2006; Nakashika *et al.*, 2016; Kinnunen *et al.*, 2017) et la conversion de voix chantée (Villavicencio & Bonada, 2010; Villavicencio & Kenmochi, 2011; Doi *et al.*, 2012; Kobayashi & Toda, 2014). La conversion vocale a un large éventail d'applications : du divertissement (parler avec la voix d'une autre personne, par exemple via des applications mobiles), créative (reconstruire la voix de personnalités), et médicale (≪réparation vocale≫ pour les personnes présentant un handicap vocal). Considérant la conversion de voix chantée, les applications créatives sont importantes dans l'industrie musicale : du karaoké à la production musicale jusqu'aux chanteurs virtuels, en contrôlant l'identité d'un chanteur réel ou artificiel (Villavicencio & Bonada, 2010; Kenmochi, 2010).
Les algorithmes de conversion d'identité vocale reposent principalement sur la conversion spectrale : la conversion du signal vocal est limitée à la conversion du timbre représenté au moyen d'enveloppes spectrales. La conversion de la voix consiste alors à apprendre une fonction de conversion entre l'espace acoustique d'une voix source et d'une voix cible. La fonction de conversion est généralement modélisée par des modèles statistiques, historiquement avec les modèles de mélange Gaussiens (GMM, (Stylianou *et al.*, 1998)) et plus récemment avec des réseaux de neurones (Desai *et al.*, 2009; Sun *et al.*, 2015). La fonction de conversion est alors apprise à partir d'une base de données pré-alignée (dite ≪parallèle≫) dans laquelle les voix source et cible ont prononcé le même ensemble de phrases, de sorte qu'une correspondance directe entre les trames des voix source et cible puisse être établie pour l'apprentissage.
Ce paradigme de conversion de la voix présente cependant des limites importantes et bien connues : les effets de sur-apprentissage et de moyennage relatifs à la modélisation statistique (Toda *et al.*, 2007) qui conduit à une dégradation de la voix convertie, et la nécessité de construire des bases de données parallèles extrêmement restrictives et non souhaitées pour des applications réelles. Pour répondre à ces limitations, des algorithmes de conversion à partir d'unités - ou d'exemples réels - ont été récemment proposés (Sündermann *et al.*, 2006; Wu *et al.*, 2013; Aihara *et al.*, 2014; Jin *et al.*, 2016). Tout d'abord, la conversion vocale à partir de sélection et de concaténation d'unités spectrales présente l'avantage de préserver les caractéristiques et la dynamique d'origine de la voix cible dans la mesure où elle repose uniquement sur l'utilisation d'unités vocales réelles. Deuxièmement, la segmentation de la parole en unités linguistiques telles que les phonèmes (voir, par exemple, la conversion vocale entre des langues différentes (Sündermann *et al.*, 2006; Nakashika *et al.*, 2016; Kinnunen *et al.*, 2017)) offre la possibilité d'utiliser des bases de données ≪à la volée≫ (dite ≪non-parallèles≫) des voix source et cible. Néanmoins, le choix de la fonction de coût utilisée pour la sélection d'unités constitue un défi important de la conversion par sélection d'unités : la mesure de la distorsion spectrale entre les voix source et cible (Sündermann *et al.*, 2006) peut être efficace dans une certaine mesure, notamment lorsque peu de données de la voix cible sont disponibles. En revanche, elle n'en demeure pas moins limitée par définition (Sündermann *et al.*, 2007) : en effet, la voix la plus proche spectralement de la voix source ne serait, à la limite, qu'elle-même. Cette limitation démontre la nécessité de considérer d'autres facteurs pour la conversion par sélection d'unités.

Les algorithmes de conversion de la voix chantée reposent sur le même paradigme que ceux utilisés pour la voix parlée (Villavicencio & Bonada, 2010; Villavicencio & Kenmochi, 2011; Doi *et al.*, 2012; Kobayashi & Toda, 2014), sans vraiment tenir compte des spécificités de la voix chantée telles que le registre étendu de hauteur, d'intensité et de durée de la voix chantée par comparaison à la voix parlée. Cet article propose un algorithme de conversion de la voix par sélection et concaténation d'unités spectrales avec un focus sur la conversion de la voix chantée. La principale contribution de l'article repose sur l'observation que les enveloppes spectrales d'un chanteur varient en fonction de plusieurs facteurs : linguistique (phonèmes), et musicale (hauteur, intensité et durée) (voir par exemple (Joliveau *et al.*, 2005) sur la voix chantée). En conséquence, la correspondance des unités spectrales d'un chanteur source et d'un chanteur cible doit être recherchée autour d'un contexte musical similaire (par exemple, mêmes hauteurs, intensités, et durées). L'algorithme de conversion vocale proposé est basé sur la sélection d'unités de phonèmes, capitalisant les avantages de la conversion par sélection d'unités et de la conversion de voix non-parallèle. Pour intégrer les spécificités de la voix chantée dans l'algorithme de sélection d'unités, une fonction de coût multiple est établie à partir de : la distorsion spectrale entre les chanteurs source et cible, et les informations musicales telles que les différences de hauteur, d'intensité et de durée entre les chanteurs source et cible. Cette fonction de coût multiple est définie afin de guider la sélection vers des enveloppes spectrales provenant de contextes musicaux similaires, et ainsi d'augmenter la correspondance entre les espaces acoustiques des chanteurs source et cible relativement à ces contextes. En d'autres termes et par analogie, les enveloppes spectrales sélectionnées de la voix cible doivent représenter pour la voix cible ce que les enveloppes spectrales de la voix source représentent pour la voix source.
L'article est organisé de la manière suivante : la conversion de voix par sélection d'unité est présentée dans la Section 2 et l'algorithme de sélection d'unités proposé avec la fonction de coût multi-cible est détaillé dans la Section 2.1. Une expérience perceptive est rapportée dans la Section 3 pour évaluer la conversion multi-cible et quelques variantes sur une tâche de conversion appliquée sur la sortie d'un synthétiseur de voix chantée.

# 2 Conversion de l'Identité de la Voix Chantée

Cette section présente un rapide aperçu des algorithmes de conversion vocale basés sur la sélection d'unités d'enveloppes spectrales, suivi d'une description détaillée de l'algorithme de conversion vocale concaténative non parallèle (coVoC) (voir (Lorenzo-Trueba *et al.*, 2018)). L'algorithme est basé sur une bases de données non-parallèle des chanteurs source et cible. Les bases de données comprennent des analyses des signaux vocaux tels que la fréquence fondamentale F0, l'intensité, les enveloppes spectrales calculées en échelle de fréquence Mel, et des transcriptions phonétiques alignées sur le signal vocal. Les principales caractéristiques de l'algorithme coVoC sont : l'exploitation des unités de phonèmes, la normalisation des différences spectrales des voix source et cible, et la nouvelle fonction de coût multi-cible qui intègre les connaissances musicales sur les chanteurs source et cible.

## 2.1 Conversion de l'identité par sélection d'unités spectrales

Le problème général de la conversion de l'identité vocale consiste à déterminer la séquence la plus vraisemblable des enveloppes spectrales de la voix cible $\mathbf{x}^{tgt} = [\mathbf{x}_1^{tgt}, \ldots, \mathbf{x}_T^{tgt}]$ à partir de la séquence

observée des enveloppes spectrales de la voix source $\mathbf{x}^{src} = [\mathbf{x}_1^{src}, \ldots, \mathbf{x}_T^{src}]$ :

$$\widehat{\mathbf{x}}^{tgt} = \underset{\mathbf{x}^{tgt}}{\operatorname{argmax}}\ p(\mathbf{x}^{tgt}|\mathbf{x}^{src}) \tag{1}$$

La solution pour la conversion par sélection d'unités (Sündermann *et al.*, 2006) est obtenue classiquement par minimisation d'une fonction de coût comprenant un coût cible défini pour chaque trame et un coût de concaténation défini entre des trames successives :

$$p(\mathbf{x}^{tgt}|\mathbf{x}^{src}) = \sum_{t=1}^{T} \mathcal{C}^t(\mathbf{x}_t^{tgt}, \mathbf{x}_t^{src}) + \mathcal{C}^c(\mathbf{x}_t^{tgt}, \mathbf{x}_{t-1}^{tgt}) \tag{2}$$

où $\mathcal{C}^t(\mathbf{x}_t^{tgt}, \mathbf{x}_t^{src})$ est la distorsion spectrale entre les trames source et cible, et $\mathcal{C}^c(\mathbf{x}_t^{tgt}, et\mathbf{x}_{t-1}^{tgt})$ est la distance euclidienne entre les trames cibles sélectionnées. La séquence $\mathbf{x}^{tgt}$ la plus vraisemblable est alors déterminée en utilisant un algorithme de Viterbi.

## 2.2 Conversion à partir d'unités phonétiques

L'algorithme de conversion présenté ci-dessus est tout d'abord étendu pour pouvoir exploiter des unités longues commes les phonèmes. L'avantage est d'une part de contraindre la conversion vocale à partir des informations linguistiques disponibles, et d'autre part de préserver la dynamique naturelle de la voix sur l'échelle des phonèmes. Pour ce faire, la sélection d'unités est reformulée en termes d'unités phonétiques :

$$\widehat{\mathbf{u}}^{tgt} = \underset{\mathbf{u}^{tgt}}{\operatorname{argmax}}\ p(\mathbf{u}^{tgt}|\mathbf{u}^{src}) \tag{3}$$

où $\mathbf{u} = [\mathbf{u}_1, \ldots, \mathbf{u}_N]$ est une partition de la séquence de trames $\mathbf{x}$ en $N$ unités phonétiques, et $\mathbf{u}_n = [\mathbf{x}_{n_1}, \ldots, \mathbf{x}_{L_n}]$ est la séquence de trames de longueur $L_n$ correspondant à l'unité $n$ et au label phonétique $l_n$.
La sélection est réalisée sous la contrainte que la séquence des étiquettes des phonèmes cibles est la même que la séquence des étiquettes de phonèmes sources, de sorte que :

$$l(\mathbf{u}_n^{tgt}) = l(\mathbf{u}_n^{src}) = l_n, \qquad \forall n \in [1, N] \tag{4}$$

où $l(\mathbf{u}_n) = l_n$ est le label du $n$-ième phonème de la séquence. En d'autres termes, chaque unité cible $\mathbf{u}_n^{tgt}$ doit être sélectionnée parmi l'ensemble des unités cibles candidates correspondant au label phonétique $l_n$, et référencée à partir de maintenant par $\mathcal{U}_{l_n}^{tgt}$.
L'optimisation est similaire à celle décrite dans la section précédente, à l'exception que la fonction de coût est définie sur les $N$ unités au lieu des $T$ trames. La fonction de coût est alors définie par comparaison d'unités phonétiques en place de la comparaison de trames.

## 2.3 Fonction de coût spectral

La distorsion spectrale $\mathcal{C}_{spec}^t$ entre les unités source et cible $\mathbf{u}_n^{src}$ et $\mathbf{u}_j^{tgt} \in \mathcal{U}_{l_n}^{tgt}$ de longeurs différentes est calculée en utilisant la déformation temporelle dynamique (DTW) comme :

$$\mathcal{C}_{spec}^t \quad = \quad D(A(\mathbf{u}_j^{tgt}, \mathbf{u}_n^{src}), \mathbf{u}_n^{src}) \tag{5}$$

où $D(.,.)$ est la distance euclidienne et $A(\mathbf{u}_j^{tgt}, \mathbf{u}_n^{src})$ la séquence de l'unité cible déformée temporellement $\mathbf{u}_j^{tgt}$ et alignée avec la séquence de l'unité source $\mathbf{u}_n^{src}$.

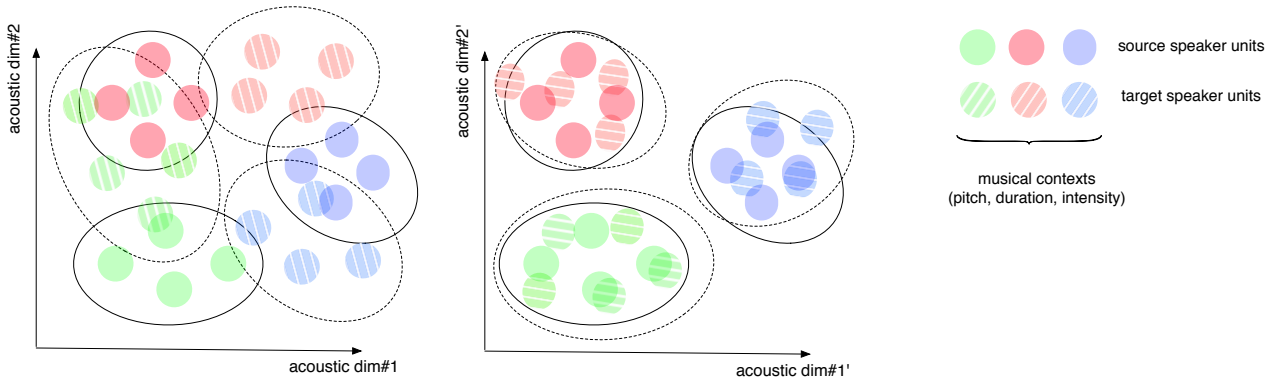



FIGURE 1 – Illustration de la fonction de côut multi-cible pour aligner les unités des voix source et cible dans l'espace musical. Chaque point représente une unité phonétique et sa séquence d'enveloppe spectrale correspondante.

Bien que la distorsion spectrale se soit montrée efficace en première approximation pour la conversion de la voix (Sündermann *et al.*, 2006), elle reste extrêmement limitée, en particulier en termes de similarité entre le locuteur source converti et le locuteur cible (Sündermann *et al.*, 2007).

En effet, la distorsion spectrale ne mesure que la similarité absolue entre les unités source et cible. Or, pour la conversion d'identité vocale, le coût entre une unité source et une unité cible doit être définie de manière à mesurer la similarité des unités *relativement* à chaque locuteur : la sélection doit être opérée de sorte à ce que l'enveloppe cible sélectionnée doit représenter pour la voix cible ce que l'enveloppe source représente pour la voix source. Pour ce faire, il est donc nécessaire d'aligner au préalable les distributions des unités source et cible pour pouvoir mesurer leur similarité de manière cohérente et relative à chaque chanteur. Ceci peut être obtenu en transformant l'espace acoustique des voix source et cible : soit en utilisant un pré-traitement de normalisation, soit en définissant une fonction de coût cible qui tient compte des multiples facteurs affectant la variabilité acoustique des chanteurs. Nous présentons ci-dessous des solutions pour ces deux types de transformation.
Une première source de différence entre l'espace acoustique des chanteurs source et cible réside dans la différence intrinsèque entre les chanteurs et entre les conditions d'enregistrement. Pour supprimer ces différences, un filtre moyen $F_l$ est créé pour chaque phonème $l$, défini comme le rapport entre les moyennes des enveloppes source et cible observées pour ce phonème. La version normalisée de la distorsion spectrale s'écrit alors :

$$\mathcal{C}^t_{spec} \quad = \quad D(A(\mathbf{u}^{tgt}_j, F\mathbf{u}^{src}_n), F\mathbf{u}^{src}_n) \tag{6}$$

Afin d'augmenter la pertinence perceptive de la mesure de distorsion, la représentation de l'enveloppe spectrale utilisée pour l'algorithme DTW et pour le calcul de la distorsion spectrale utilise une échelle de fréquence Mel.

Le coût de concaténation correspondant $\mathcal{C}^c_{spec}$ entre les unités déformées temporellement du chanteur cible est alors défini comme suit :

$$\mathcal{C}^c_{spec} \quad = \quad D(\mathbf{u}^{tgt}_{i,r}, \mathbf{u}^{tgt}_{j,l}) \tag{7}$$

où $\mathbf{u}^{tgt}_{i,r}$ et $\mathbf{u}^{tgt}_{j,l}$ représentent les trames droite (de fin) de l'unité $\mathbf{u}^{tgt}_i$ et gauche (de début) de l'unité $\mathbf{u}^{tgt}_j$.

## 2.4 Fonction de coût multi-cible

Une deuxième source de différence réside dans le fait que l'enveloppe spectrale d'un locuteur varie en fonction de multiples facteurs, tels que la hauteur, l'intensité et la durée (voir par exemple (Joliveau *et al.*, 2005) sur la voix chantée). Pour compenser ces différences, la comparaison des unités spectrales d'un chanteur source et d'un chanteur cible doit être mesurée de manière relative au contexte musical dans lequel elles sont observées. Pour ce faire, une fonction de coût multi-cibles est proposée afin de prendre en compte ces facteurs, avec comme motivation principale de guider la sélection des unités spectrales vers des unités provenant d'un contexte musical similaire. En conséquence, la fonction de coût multi-cible proposée est écrite de la manière suivante :

$$\mathcal{C}^t = \sum_{factor} w_{factor} \mathcal{C}^t_{factor} \tag{8}$$

où $\mathcal{C}^t_{factor}$ représentent les fonctions de coûts partiels mesurant la similarité entre la source et la cible en fonction de l'un des facteurs, et $w_{factor}$ les poids correspondant attribués aux facteurs. Comme mentionné dans (Taylor, 2006), la définition d'une fonction de coût cible multiple peut être interprétée comme une projection des unités d'origine dans un espace où chaque facteur est représenté par sa propre dimension dont la métrique est définie par le coût partiel $\mathcal{C}^t_{factor}$, avec le facteur de mise à l'échelle $w_{factor}$. Dans cet article, les facteurs considérés sont les facteurs spectraux, la hauteur, l'intensité et la durée. La fonction de coût multiple peut alors être interprétée comme une projection des unités spectrales source et cible afin de les aligner *relativement* aux facteurs musicaux, comme le montre la Figure 1. Concrètement, l'effet de cette projection est de rapprocher les enveloppes spectrales des chanteurs source et cible issues d'un contexte musical similaire, définies par la hauteur, l'intensité et la durée des unités correspondantes.
La fonction des coûts partiels est définie en utilisant la distance euclidienne entre les valeurs moyennes entre les caractéristiques normalisées mesurées sur les unités source et cible. Par exemple :

$$\mathcal{C}^t_{F0} = D(\overline{\mathbf{u}^{src}_i}(\mathbf{F0}^{src}_{norm}), \overline{\mathbf{u}^{tgt}_j}(\mathbf{F0}^{tgt}_{norm})) \tag{9}$$

avec $\mathbf{F0}_{norm}$ la séquence de F0 normalisée du chanteur, et $\overline{\mathbf{u}_i}(\mathbf{F0}_{norm})$ la F0 normalisée moyenne de la $i$-ème unité. Les autres fonctions de coût partiel $\mathcal{C}^t_{int}$ et $\mathcal{C}^t_{dur}$ sont définies de la même manière. La conversion vocale multi-cible a été utilisée avec les poids suivants : $w_{spec} = 15$, $w_{F0} = 10$ et $w_{dur} = 5$, qui ont été choisi empiriquement à partir d'essais informels. Le facteur d'intensité n'a pas été utilisé puisque la base de données utilisées pour la synthèse de chant avait une dynamique presque constante sur chaque unité de diphone (Ardaillon, 2017).

# 3 Expérience

## 3.1 Matériel

L'algorithme de sélection d'unités multi-cibles proposé a été évalué dans une expérience de perception sur la conversion de la voix chantée. Le chanteur cible est un chanteur français qui a enregistré huit chansons du chanteur français Jacques Brel. Les enregistrements ont été réalisés dans des conditions professionnelles (studio d'enregistrement, ingénieur du son), et numérisés avec une fréquence d'échantillonnage de 48.000 Hz et avec un encodage de 16 bits par échantillon. La voix chantée de la source a été créée en utilisant un synthétiseur de voix chantée (Ardaillon, 2017) à partir

des partitions musicales des chansons. Le but de l'utilisation d'un synthétiseur vocal est d'évaluer la conversion de la voix dans des conditions contrôlées, en garantissant le respect de la partition musicale et la connaissance précise des labels et des limites des phonèmes. En outre, l'utilisation d'une voix de synthèse constitue une preuve de concept que l'identité d'une voix de synthèse peut être contrôlée efficacement par conversion de l'identité vocale (Villavicencio & Bonada, 2010).

## 3.2 Banc d'essai des algorithmes de conversion

Un banc d'essai de systèmes de conversion à partir de sélection d'unités a été élaboré pour comparaison : d'une part, une fonction de coût basée sur la distorsion spectrale seule, et d'autre part la fonction de coût multi-cibles proposée. L'un des avantages de l'algorithme de conversion vocale proposé réside dans la possibilité de restreindre la conversion à un sous-ensemble de phonèmes et de garder le reste des phonèmes inchangés. En conséquence, cet article évalue également la conversion vocale chantée obtenue en convertissant tous les phonèmes, ou en convertissant seulement les voyelles du chanteur source. Ceci est basé sur l'hypothèse que l'identité est majoritairement portée par les voyelles, qui sont en outre plus stables et beaucoup plus longues que les consonnes dans la voix chantée. Pour résumer, les échantillons utilisés pour l'expérience sont : chant synthétisé source (S), chanteur cible (T), conversion de tous les phonèmes par distorsion spectrale (SD), conversion de tous les phonèmes avec la fonction de coût multi-cible (MO), et la même conversion vocale en convertissant seulement les voyelles (respectivement, SD_VOW et MO_VOW).

## 3.3 Configurations expérimentales

L'expérience a consisté dans le jugement par des auditeurs d'échantillons vocaux chantés, basés sur la similarité avec le chanteur cible et le caractère naturel du chanteur, comme utilisé pour la compétition de conversion d'identité vocale de 2016 (Toda *et al.*, 2016). Pour ce faire, quatre chansons ont été sélectionnées pour l'expérience parmi les huit disponibles, et les deux premières phrases de ces quatre chansons ont été utilisées pour la conversion. Pour une chanson donnée, la conversion de la voix a été effectuée en utilisant les morceaux disponibles restants, soit les sept chansons restantes. Pendant l'expérience, le participant devait choisir entre l'une des quatre chansons à évaluer. Ensuite, les échantillons vocaux chantés (originaux, synthétisés et convertis) étaient présentés dans un ordre aléatoire, avec toujours la possibilité d'écouter le vrai chanteur cible, et le participant devait évaluer le naturel de l'échantillon et la similarité avec le chanteur cible. L'expérience a été menée en ligne, encourageant l'utilisation d'écouteurs et de casque audio dans un environnement silencieux. Vingt personnes ont participé à l'expérience. Chaque personne a évalué les conversions d'une seule chanson, soit 2 phrases fois 6 versions (incluant le chant synthétisé S et la voix cible T ).

## 3.4 Résultats et Discussion

La Figure 2 présente les scores obtenus par la voix chantée synthétisée, le chanteur cible et les algorithmes de conversion vocale. La voix chantée synthétisée a un naturel acceptable (3,3) mais la plus faible similarité avec la voix cible (2,1). Tous les algorithmes de conversion ont une similarité significativement plus élevée, ce qui vient malheureusement avec une dégradation du signal de chant converti. Par comparaison des algorithmes VC : l'algorithme multi-cible proposé a une similarité significativement plus élevée avec le chanteur cible (3,2) que la conversion basée sur la distorsion spectrale (2,7), et avec un naturel comparable (respectivement, 2,0 et 1,9). Cela montre que la prise en compte du contexte musical dans la recherche d'enveloppes spectrales améliore la

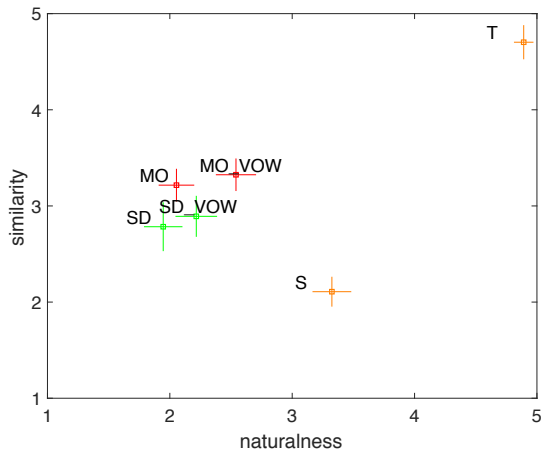



FIGURE 2 – Scores de conversion : score moyen et intervalles de confiance à 95%.

sélection d'enveloppes spectrales adéquates, conduisant à un gain de naturel et de similarité. Enfin, la préservation des consonnes conduit à un naturel significativement plus élevé (respectivement, 2,5 et 2,2) qui vient également avec une augmentation de la similarité (respectivement, 3,3 et 2,9). Bien que le gain de naturel soit clairement attendu, le gain de similarité suggère que tous les phonèmes n'ont pas la même importance dans la conversion de la voix. En particulier, les voyelles peuvent transmettre plus d'informations sur l'identité du chanteur que les consonnes. D'autre part, ce résultat suggère que les deux dimensions de jugement de la conversion ne sont clairement pas orthogonales : la dégradation de la conversion affecte directement le jugement de la similarité à la voix cible.

# 4 Conclusion

Cet article a présenté un algorithme de sélection d'unités multi-cibles pour la conversion non-parallèle de la voix chantée. L'idée principale est que la sélection des enveloppes spectrales doit être faite pour que les enveloppes spectrales sélectionnées du chanteur cible soient non seulement similaires mais aussi issues d'un même contexte linguistique (phonèmes) et musical (hauteur, intensité, durée) que celles du chanteur source. Une expérience perceptive a été menée pour convertir un synthétiseur vocal en un célèbre chanteur français. La conversion vocale multi-cible proposée a été jugée sensiblement plus similaire au chanteur cible par rapport à un algorithme classique de sélection d'unités. Les recherches futures porteront sur l'apprentissage des enveloppes spectrales en fonction des facteurs musicaux et leur exploitation en conversion de la voix chantée par sélection d'unités.

# Références

AIHARA R., NAKASHIKA T., TAKIGUCHI T. & ARIKI Y. (2014). Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary. In *International Conference on Audio, Speech, and Signal Processing (ICASSP)*, p. 7944–7948.

ARDAILLON L. (2017). *Synthesis and expressive transformation of singing voice*. PhD thesis, Ircam-Upmc, Paris, France.

DESAI S., RAGHAVENDRA E. V., YEGNANARAYANA B., BLACK A. W. & PRAHALLAD K. (2009). Voice conversion using artificial neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

DOI H., TODA T., NAKANO T., GOTO M. & NAKAMURA S. (2012). Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system. In *Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC)*.

JIN Z., FINKELSTEIN A., DIVERDI S., LU J. & MYSORE G. J. (2016). CUTE : A concatenative method for voice conversion using exemplar-based unit selection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

JOLIVEAU E., SMITH J. & WOLFE J. (2005). Vocal tract resonances in singing : The soprano voice. *Journal of the Acoustical Society of America*, **116**(4), 2434–2439.

KENMOCHI H. (2010). VOCALOID and Hatsune Miku phenomenon in Japan. In *Intersinging, Interdisciplinary Workshop on Singing Voice*.

KINNUNEN T., JUVELA L., ALKU P. & YAMAGISHI J. (2017). Non-parallel voice conversion using i-vector plda : towards unifying speaker verification and transformation. In *International Conference on Audio, Speech, and Signal Processing (ICASSP)*.

KOBAYASHI K. & TODA T. (2014). Statistical singing voice conversion with direct waveform modification based on the spectrum differential. In *Interspeech*, p. 2514–2518.

LORENZO-TRUEBA J., YAMAGISHI J., TODA T., SAITO D., VILLAVICENCIO F., KINNUNEN T. & LING Z. (2018). The voice conversion challenge 2018 : Promoting development of parallel and nonparallel methods. In *Speaker Odyssey*.

NAKASHIKA T., TAKIGUCHI T. & MINAMI Y. (2016). Non-parallel training in voice conversion using an adaptive restricted boltzmann machine. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24**(11), 2032 – 2045.

STYLIANOU Y., CAPPÉ O. & MOULINES E. (1998). Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, **6**(2), 131–142.

SUN L., KANG S., LI K. & MENG H. (2015). Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

SÜNDERMANN D., HÖGE H., BONAFONTE A., NEY H., BLACK A. & NARAYANAN S. (2006). Text-independent voice conversion based on unit selection. In *International Conference on Audio, Speech, and Signal Processing (ICASSP)*, p. 1173–1176.

SÜNDERMANN D., SMREKAR J., HÖGE H., BONAFONTE A. & NEY H. (2007). The speech alignment paradox. In *International Workshop on Advances in Speech Technology (AST)*.

TAYLOR P. (2006). The target cost formulation in unit selection speech synthesis. In *Interspeech*.

TODA T., BLACK A. W. & TOKUDA K. (2007). Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, **15**(8), 2222–2235.

TODA T., CHEN L.-H., SAITO D., VILLAVICENCIO F., WESTER M., WU Z. & YAMAGISHI J. (2016). The voice conversion challenge 2016. In *Interspeech*.

VILLAVICENCIO F. & BONADA J. (2010). Applying voice conversion to concatenative singing-voice synthesis. In *Interspeech*, p. 803–806.

VILLAVICENCIO F. & KENMOCHI H. (2011). Non-parallel singing-voice conversion by phoneme-based mapping and covariance approximation. In *DAFx*, p. 241–244.

WU Z., VIRTANEN T., KINNUNEN T., CHNG E. S. & LI H. (2013). Exemplar-based unit selection for voice conversion utilizing temporal information. In *Interspeech*.

# Mesure de similarité fondée sur des réseaux de neurones siamois pour le doublage de voix

Adrien Gresse Richard Dufour Vincent Labatut
Mickaël Rouvier Jean-François Bonastre
LIA - Université d'Avignon (France)
prénom.nom@univ-avignon.fr

RÉSUMÉ

Le doublage vocal d'une œuvre culturelle permet sa diffusion vers une audience plus large. Le processus de sélection de voix dans une nouvelle langue, intégralement réalisé par un opérateur humain, est appelé casting vocal. Cette sélection dépasse le simple cadre de la proximité acoustique entre deux voix, intégrant de nombreux critères plus subjectifs qui peuvent être liés notamment à des choix socioculturels, émotionnels... Dans ce papier, nous proposons une approche par réseaux de neurones siamois mesurant la proximité entre la voix originale et la voix dans la langue cible, en intégrant la notion de similarité entre les voix non pas d'un point de vue purement acoustique mais également réceptif. Les premiers résultats obtenus montrent, grâce à un test d'hypothèse statistique, que des informations sont contenues dans les paramètres acoustiques pour un même personnage permettant à une voix d'être associée à une autre.

ABSTRACT

**Siamese neural networks based similarity metric for dubbing**

Dubbing aims to broadcast a multimedia document to a larger audience. The process that consists in selecting a voice in a target language is referred as voice casting and it is performed by a human. This selection is not only based on acoustic similarity between two voices. Actually, it is suppported by more subjective criteria such as emotions, sociocultural choices... In this paper we propose a siamese neural networks based approach measuring proximity between the original voice and the dubbed one. The concept of similarity we want to model does not only consider the acoustic part of a voice, also it takes into account spectators receptive concerns. We perform a statistical test to evaluate our model. Our results show that there is an information in the acoustic parameters that allows a voice to be associated with another one with respect to a particular character.

MOTS-CLÉS : casting vocal, réseaux de neurones siamois, $i$-vecteur, similarité.

KEYWORDS: voice casting, siamese neural networks, $i$-vector, similarity.

# 1 Introduction

La voix apparaît, dans de nombreuses œuvres culturelles (films, documentaires, jeux vidéos...), comme un vecteur de stimulis émotifs pour le public qui la reçoit. Dans un contexte de diffusion internationale, un doublage vocal est souvent réalisé en remplaçant la voix originale par une nouvelle voix dans une langue cible. Le processus qui permet de sélectionner, à partir d'une voix originale, une voix parmi plusieurs voix candidates dans une autre langue est appelé *casting vocal*. Il s'agit de l'objet d'étude principal de nos travaux. À l'origine, la sélection est réalisée par un opérateur humain


10.21437/JEP.2018-2

en fonction, d'une part, de la voix originale et de critères plus subjectifs pouvant être liés, par exemple, au personnage, ou rôle, interprété. En sciences humaines, le terme *réception* est utilisé pour parler des effets à long terme d'une voix perçue à un moment donné. Il ne s'agit pas ici de trouver la voix la plus semblable à celle d'origine au niveau acoustique, mais de trouver celle qui aura, dans cette nouvelle langue, un effet identique à la voix d'origine, faisant intervenir des critères socioculturels, ou autres.

Un des enjeux du travail que nous menons réside dans la notion de "similarité" entre les voix. D'une manière générale, celle-ci a déjà été étudiée à maintes reprises. Nombreux sont les papiers qui ont découlé du travail de Laver (1980), qui propose un moyen de décrire la *qualité vocale*, se comprenant comme les caractéristiques auditives qui colorent la voix d'un individu. Plusieurs travaux proposent d'évaluer le degré de similarité de la voix perçue dans un groupe de voix (McDougall, 2013; Rose, 1999; Loakes, 2006; Nolan *et al.*, 2011; Baumann & Belin, 2010), bien souvent dans le cadre d'applications juridiques. Entre autres, ces travaux montrent qu'il existe des corrélations entre certaines caractéristiques acoustiques et le fait que des voix soient perçues comme étant similaires ou non. Toutefois, il n'existe pas de méthode établie pour quantifier le degré de similarité entre deux voix. Dans le vaste domaine de la reconnaissance automatique du locuteur, des systèmes permettent indirectement de mesurer la distance entre deux identités locuteurs, et par conséquent d'évaluer la "similarité" de leur voix (Kelly *et al.*, 2016; Zhang & Tan, 2008; Lindh & Eriksson, 2010). Néanmoins, cette notion de similarité n'induit principalement qu'une ressemblance acoustique. Dans notre contexte, nous souhaitons étendre cette notion de similarité en y intégrant d'autres critères de nature plus humaine, qui guident le choix de l'opérateur de casting vocal. Nous pensons aux éléments amenés notamment par le jeu de l'acteur tels que les inflexions de voix utilisées pour mieux faire ressortir les traits du personnage incarné. Récemment, certains travaux ont commencé à explorer cette dimension (Obin *et al.*, 2014; Obin & Roebel, 2016; Gresse *et al.*, 2017) qui offrent une comparaison entre l'utilisation d'un système de reconnaissance automatique du locuteur et d'un classifieur multimodal de critères para-linguistiques.

Nos travaux s'inscrivent dans le cadre de la recommandation automatique de voix pour des œuvres culturelles. Dans cette optique, nous proposons d'explorer la dimension caractéristique du personnage perçue au travers de la voix dans le cadre d'un jeu vidéo. Il s'agit d'une approche inédite pour la mesure de la similarité de la voix dans un contexte multilingue. Nous nous limitons pour le cas présent à deux langues (anglais et français). À la différence de (Gresse *et al.*, 2017), où nous avions proposé une approche $i$-vecteur/PLDA inspirée de la reconnaissance du locuteur, nous explorons ici l'utilisation de réseaux de neurones siamois. En effet, nous avons observé que la PLDA a tendance à se focaliser sur l'identité du locuteur. De plus, notre méthode pouvait s'apparenter à un mapping des locuteurs dans la langue originale vers les locuteurs de la langue source, plus qu'à une estimation pure de la similarité entre les deux. Notre intuition est que les réseaux de neurones siamois devraient être mieux adaptés à la notion de similarité telle que définie dans le casting vocal. Nous nous concentrons ici sur la mesure de la similarité plus que sur la classification des voix. Nous utilisons un test statistique pour vérifier si le modèle est capable d'abstraire cette notion de similarité.

Le reste de cet article est organisé comme suit. Dans la Section 2 nous détaillons notre approche. La méthode et le protocole expérimental sont décrits dans la Section 3 avant de présenter nos résultats dans la Section 4. Enfin, nos conclusions et nos perspectives futures sont énoncées dans la Section 5.

# 2 Approche proposée

La Figure 1 illustre de manière simplifiée notre système automatique de casting vocal. Ce dernier accepte deux entrées qui correspondent à deux extraits de voix, et une sortie qui représente le score

de similarité entre celles-ci, estimant leur degré de proximité. Pour le casting vocal, le système doit prédire dans quelle mesure la voix d'une langue *cible* peut être utilisée pour le doublage de la voix dans une langue *source*, en dépassant la simple ressemblance acoustique. Au cœur du système se trouve notre modèle de similarité appris sur un ensemble de voix.

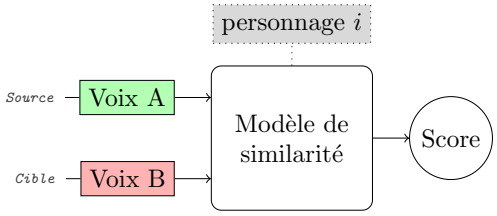



FIGURE 1 – Présentation simplifiée du système automatique de casting vocal.

Dans la Section 2.1, nous présentons les concepts et notre motivation quant à l'utilisation des réseaux de neurones siamois, constituant l'originalité de ce travail. Les données en entrée de ces réseaux sont représentées au moyen de $i$-vecteurs, que nous présentons succinctement dans la Section 2.2.

## 2.1 Réseaux de neurones siamois

De manière intuitive, les architectures siamoises nous offrent un moyen d'apprendre une mesure de similarité à partir de deux entrées indépendantes qui partagent une relation abstraite de similarité. Les premiers travaux faisant utilisation de réseaux de neurones siamois font référence à (Bromley *et al.*, 1994) pour la vérification de signatures. Ce type d'architecture a la particularité de faire intervenir deux réseaux de neurones identiques qui prennent deux entrées indépendantes et qui se rejoignent finalement grâce à une fonction de pénalité (voir Figure 2). Cette fonction se base sur une métrique (ici une distance) calculée à partir des représentations de plus haut-niveau des deux réseaux. À noter que les deux réseaux qui interviennent dans ce type d'architecture partagent les mêmes paramètres.

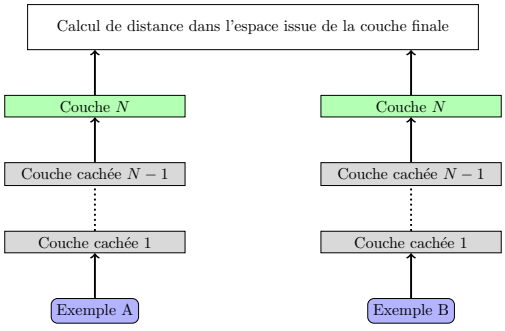



FIGURE 2 – Exemple de réseaux de neurones siamois.

Dans notre travail, nous avons mis en place une architecture siamoise semblable à (Chopra *et al.*, 2005; Hadsell *et al.*, 2006). Elle nous assure deux choses (Koch *et al.*, 2015) :

— Du fait du partage des paramètres, des entrées fortement similaires ne peuvent pas être projetées à des endroits différents dans l'espace de représentation latent et inversement, une paire d'entrées différentes ne peut être projetée par les réseaux siamois à des endroits proches.
— Aucune distinction n'est faite par la fonction de pénalité quant à l'ordre des entrées constituant la paire traitée (*i.e.* la fonction de similarité est symétrique).

Dans (Chopra *et al.*, 2005) les auteurs utilisent une fonction de pénalité s'appuyant sur une mesure de l'énergie définie comme $E_W(I_1, I_2) = \|G_W(I_1) - G_W(I_2)\|_2$, où $I_1$ et $I_2$ correspondent aux entrées et $G$ représente une fonction de projection depuis l'espace des entrées vers un nouvel espace de représentation. En jouant sur les paramètres $W$, il faut donc minimiser l'énergie lorsque les deux entrées $I_1$ et $I_2$ sont similaires mais aussi s'assurer que $E_W$ est grande pour des entrées différentes. À juste titre, la fonction de pénalité est qualifiée de contrastive. Soit une variable binaire notée $T$ telle que $T = 0$ lorsque les entrées sont similaires et $T = 1$ dans le cas contraire. On considère une constante notée $m$ positive que l'on peut interpréter comme une marge. La fonction de pénalité est définie par l'équation suivante :

$$L(I_1, I_2, T) = (1 - T) \times (E_W(I_1, I_2))^2 + T \times max\{0, m - E_W(I_1, I_2)\}^2$$

Dans les travaux présentés, nous avons utilisé deux réseaux Perceptron multicouches (MLP) contenant 2 couches cachées de 1 000 unités plus une couche finale constituées de 500 unités, combinées à une fonction tangente hyperbolique. Les deux réseaux calculent la même fonction $G_W$. Enfin nous avons utilisé un distance de Manhattan pour le calcul du l'énergie.

## 2.2 Représentation des données par des $i$-vecteurs

La question du choix de la représentation des segments audio pour une mesure de similarité de voix de doublage apparaît comme une problématique en elle-même. De manière générale, le choix de la représentation des données en entrée (ici des segments audio) a une influence non-négligeable sur les performances finales des systèmes : c'est aussi le cas dans notre contexte de casting vocal. Ainsi, la contrainte principale concerne la variabilité de la durée des séquences audio. Afin de pouvoir représenter des séquences de durée variables par un vecteur de taille fixe, nous avons choisi de représenter ces données au moyen de $i$-vecteurs.

Les $i$-vecteurs ont été initialement présentés dans le domaine de la vérification du locuteur (Dehak *et al.*, 2011) et ont depuis montré leur robustesse. Ils contiennent, entre autres, les caractéristiques propres au locuteur mais également des informations liées au canal de transmission ou au contenu phonétique du segment audio. Cette représentation est extraite à partir de séquences pouvant être de tailles différentes. Plus généralement on dit que le $i$-vecteur est une représentation compacte de la séquence de paramètres acoustiques extraite à partir d'un segment de voix.

# 3 Protocole Expérimental

Dans cette section, nous détaillons les données utilisées pour notre problème de casting vocal (Section 3.1). Puis nous définissons la manière dont nos expériences seront menées (Section 3.3 et Section 3.2). Enfin, nous proposons un protocole d'évaluation de notre approche (Section 3.4).

## 3.1 Données

Nos expériences sont réalisées sur les données issues du jeu vidéo *Mass Effect 3*. Nous avons extrait les interactions vocales des différents personnages du jeu dans leurs versions originales (*i.e.* en anglais) et doublées (*i.e.* en français). Notre objectif est d'apprendre à réaliser – à l'instar de l'opérateur de casting vocal – l'appariement des voix originales et doublées de manière automatique. L'apprentissage automatique se base sur des paires de voix dont l'une d'entre elles appartient à l'ensemble des segments de voix originales (ici anglais) et la deuxième appartient quant à elle à l'ensemble des segments de voix doublées (ici français). Nous avons donc un ensemble contenant

toutes les paires réalisables $\mathcal{L} = \{(x_i, y_j)\}$ avec $x_i \in X$ l'ensemble des segments originaux en anglais et $y_j \in Y$ l'ensemble des segments en français. Nous notons au passage que les ensembles $X$ et $Y$ sont bijectifs du fait que chaque segment de la version originale possède exactement un équivalent dans la version française.

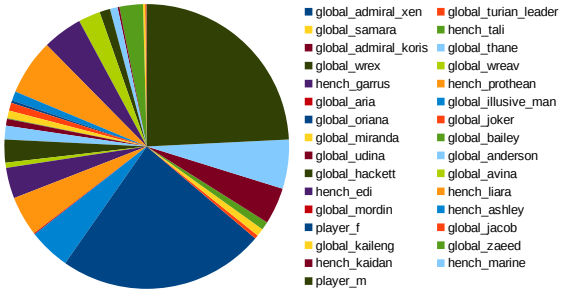



FIGURE 3 – Répartition des segments par nom de personnage.

Nous pouvons définir un personnage $k$ comme un couple de locuteurs $\mathcal{P}_k = (S^k_{en}, S^k_{fr})$ où $S^k_{en}$ représente l'ensemble des segments du locuteur correspondant au personnage $k$ en anglais (en) et $S^k_{fr}$ représentant les segments du locuteur doublant le même personnage $k$ en français (fr). Il est important de noter que nous avons veillé à ce qu'aucun locuteur ne soit associé à plus d'un personnage, dans le but d'éviter tout biais à ce niveau-là. Nous soulignons également que le nombre total de segments de voix n'est pas équitablement réparti entre les différents personnages comme on peut le voir dans la Figure 3. Le protocole expérimental a été défini afin de réduire au mieux ce biais.

Chaque ensemble de segments (*en* et *fr*) compte 10 000 segments audio de haute qualité (enregistrés en studio) pour un corpus contenant un total approximatif de 7,5 heures de dialogues dans chaque langue. Les segments ont en moyenne une durée de 3,5 secondes, et comme il a été dit plus haut, le nombre de segments par personnage est relatif à son importance dans le jeu. Ainsi, nous avons en moyenne, pour les deux langues, 12 minutes de dialogue par personnage, avec un écart-type de 20 minutes. Autre élément important : nous avons un total de 31 personnages différents associés chacun à un acteur différent (tous personnages confondus et toutes langues confondues) pour le doublage, soit 62 identités locuteurs représentées dans nos données.

## 3.2 Traitement des séquences

Le signal audio a été transformé en vecteurs de caractéristiques de 60 dimensions, avec 20 paramètres MFCC incluant l'énergie auxquels s'ajoutent les 20 dérivées du premier ordre ($\Delta$) et 20 du second ($\Delta\Delta$). Les paramètres sont calculés sur des fénêtres de 20 ms avec un décalage de 10 ms. Nous avons appliqué une normalisation sur la moyenne cepstrale et supprimé les trames de faible énergie, qui correspondent principalement à du silence. Nous avons ensuite entraîné un modèle du monde (UBM) de 2 048 composantes à partir des vecteurs de caractéristique et une matrice de variabilité totale $T$ de rang 400 qui nous permet d'extraire nos $i$-vecteurs. Nous avons créé deux espaces $i$-vecteurs, un pour l'anglais et l'autre pour le français. Le modèle du monde ainsi que la matrice $T$ ont été appris pour l'anglais sur NIST SRE 2004, 2005 et 2006. Pour le français, l'UBM et la matrice $T$ ont été appris sur les campagnes d'évaluation ESTER-1, ESTER-2, EPAC, ETAPE et REPERE.

## 3.3 Apprentissage

Les paires constituées de segments de voix doublant un même personnage sont dites *target*, toutes les autres *nontarget*. Étant donné que notre ensemble de paires est issu de la combinaison des deux ensembles de segments de voix, nous avons un nombre beaucoup plus grand de paires *nontarget* que de *target*. Pour palier ce biais, nous avons réalisé un équilibrage des paires utilisées pour les tests. En effet, il nous faut avoir un parfait équilibre entre ces paires pour le corpus de test qui nous sert de corpus de contrôle. En détails, 16 personnages disposent d'un nombre de segments supérieur ou égal à 95. Nous pouvons donc utiliser 4 de ces personnages pour notre corpus de test, ceux-ci étant par conséquent retirés du corpus d'apprentissage afin d'éviter tout effet de mémorisation. A travers une validation croisée, nous obtenons 4 ensembles de tests différents ainsi que 4 corpus d'apprentissages variant également dans la mesure où les personnages non-utilisés pour les tests y sont réintégrés.

Compte tenu de nos paires, composées d'un segment anglais (source) et d'un segment français (cible), et avec 4 personnages comptant 95 segments (tirés aléatoirement pour ceux en ayant un nombre plus grand) en anglais et 95 en français, nous avons un corpus de test composé de $4 \times 95 \times 95$ paires *target* et $12 \times 95 \times 95$ paires *nontarget*, que nous ramenons donc au même nombre par tirage aléatoire. Cette procédure permet d'éliminer les biais liés aux probabilités *a priori* des 2 classes ramenées à 0,5.

Nous avons également pensé à l'impact de la langue, ainsi qu'au contenu des segments de voix comme biais possibles. Nous utilisons toujours la même configuration de langue, soit un segment en anglais combiné avec un segment en français. Étant donné la sensibilité à la durée des $i$-vecteurs qui est directement reliée au contenu linguistique de chaque segment, nous veillons à éviter les paires qui associent un segment anglais à son homologue en français en mélangeant aléatoirement tous les segments au préalable. Enfin, nous avons également levé le biais potentiellement introduit par la différence de genre. Pour cela, nous effectuons nos tests sur les paires de même genre uniquement. Les paires *target* respectant de par nature cette dernière contrainte, nous réduisons donc l'ensemble des paires *nontarget* aux seules paires de même genre.

## 3.4 Évaluation

Notre évaluation se base sur les scores obtenus pour chaque paire. Il s'agit d'une distance de Manhattan calculée dans l'espace de représentation du modèle de similarité que nous avons appris. Les scores obtenus sur les paires issues du corpus de test sont regroupés en deux groupes. Les scores correspondants aux paires *target* dans l'un, ceux des paires *nontarget* dans l'autre. Dans le but d'évaluer la pertinence du modèle de similarité appris, nous réalisons un test statistique d'hypothèse : le $t$-test, ou test de *Student*. L'idée est de comparer la moyenne des scores des deux groupes. Il s'agit d'un test bilatéral où l'hypothèse nulle $H_0$ dit que les moyennes des deux groupes sont identiques. Pour étayer cette statistique il est impératif d'y ajouter des éléments descriptifs tels que la dispersion ou l'écart à la moyenne.

# 4 Résultats

Nous présentons dans la Table 1 l'ensemble des résultats obtenus au moyen du test statistique de *Student* sur les résultats des réseaux de neurones siamois. Pour chacun des 4 ensembles de tests détaillés plus haut, nous donnons la valeur du $t$-test avec la probabilité qui lui est associée. Ainsi nous avons obtenu des valeurs de $9,16$ et $11,22$ associées à une probabilité $< 0.01$ pour les ensembles de tests notés 1 et 2. Pour les ensembles notés 3 et 4, les valeurs du $t$-test sont de $-27,44$ et $-34,37$ aussi associées à une $p$-value $< 0.01$. Étant donné que le $t$-score est un ratio de la différence entre les

deux groupes et de la différence à l'intérieur des groupes, nous observons que la différence entre les groupes de paires *target* et *nontarget* est plus grande pour les tests notés 3 et 4. Cela peut signifier plusieurs choses. Le plus probable étant que, pour les personnages qui composent les ensembles de tests 1 et 2, le modèle ne soit pas parvenu à généraliser. Il est aussi envisageable que les personnages de chaque test soient déjà similaires entre eux ou au contraire plus variés. Nous entendons des personnages similaires au sens d'un même type de personnage (*e.g.* jouant le rôle de soldat). Ces résultats montrent que le système fait des confusions entre certains personnages. En effet, les deux groupes de scores étant très similaires, il y a donc confusion dès que l'on se retrouve à l'intersection des distributions des scores *target* et *nontarget*. Si l'on se réfère aux moyennes, nous observons une différence notamment sur les tests 1 et 2. En effet le score moyen des paires *target* est inférieur à celui des paires *nontarget* contrairement aux tests 3 et 4. Nos scores correspondent à une distance et devraient logiquement être en moyenne plus faibles chez les individus *target*. Or, nous observons le phénomène inverse dans les deux premiers tests, ce qui renforce un peu plus l'idée que la configuration des personnages pour ces tests là n'est pas adéquate. Nous avons également testé l'utilisation d'une distance euclidienne (Chopra *et al.*, 2005) au lieu d'une distance de Manhattan. Les résultats obtenus ne variant pas significativement, aussi nous n'avons pas jugé utile de les faire apparaître ici.

| # | NOMBRE DE PAIRES | CORPUS | |
|---|---|---|---|
| | | **TRAIN** | **TEST** |
| **1** | genre identique : | 19420084 | 72200 |
| | genre différent : | 8218842 | 0 |
| | target / nontarget (même nb.) | 13819463 | 36100 |
| | | $t$-score / $p$-value : | **9,16** | **4,96E-20** |
| | | | *target* | *nontarget* |
| | | moyenne : | 0,65 | 0,63 |
| | | écart-type : | 0,32 | 0,32 |
| **2** | genre identique : | 13168499 | 72200 |
| | genre différent : | 2535633 | 0 |
| | target / nontarget (même nb.) | 7852066 | 36100 |
| | | $t$-score / $p$-value : | **11,22** | **3,25E-29** |
| | | | *target* | *nontarget* |
| | | moyenne : | 0,56 | 0,54 |
| | | écart-type : | 0,23 | 0,24 |
| **3** | genre identique : | 9896643 | 72200 |
| | genre différent : | 3977211 | 0 |
| | target / nontarget (même nb.) | 6936927 | 36100 |
| | | $t$-score / $p$-value : | **-27,44** | **6,55E-165** |
| | | | *target* | *nontarget* |
| | | moyenne : | 0,54 | 0,6 |
| | | écart-type : | 0,28 | 0,26 |
| **4** | genre identique : | 18275296 | 72200 |
| | genre différent : | 8117988 | 0 |
| | target / nontarget (même nb.) | 13196642 | 36100 |
| | | $t$-score / $p$-value : | **-34,37** | **7,53E-257** |
| | | | *target* | *nontarget* |
| | | moyenne : | 0,59 | 0,67 |
| | | écart-type : | 0,29 | 0,32 |
| **TOTAL CUMULÉ** | | **t-test / p-value :** | **-21,48** | **2,71E-102** |
| | | | ***target*** | ***nontarget*** |
| | | **moyenne :** | **0,59** | **0,61** |
| | | **écart-type :** | **0,28** | **0,29** |

TABLE 1 – Résultats du test statistique de *Student* sur les corpus de tests.

Nous avons aussi effectué un test de *Student* sur toutes les paires de tests cumulées. Nous observons une différence significative compte tenu du $t$-score de $-21,48$ associé à une probabilité $< 0.01$. À titre de comparaison, nous avons réalisé le même test sur les scores obtenus avec la méthode présentée dans (Gresse *et al.*, 2017). Sur l'ensemble total de paires de tests, le $t$-score est de $2,70$

avec une probabilité également inférieure au seuil de rejet. En définitive, ces résultats nous amènent à la conclusion que l'approche présentée dans cet article est bien adaptée à la modélisation de la similarité de voix de doublage.

# 5 Conclusion

Le casting vocal, intégrant la problématique de la perception des voix, est un problème complexe du fait de la multitude de facteurs pouvant influencer le choix de l'opérateur. Cette complexité se ressent dès lors que l'on essaye d'automatiser une tâche qui n'est pas clairement formalisée et qui laisse place à des critères subjectifs. Dans cet article, nous avons proposé un système automatique pour essayer de nous rapprocher des choix de l'opérateur de casting. Ce système nous permet surtout d'explorer, au moyen de méthodes statistiques, cette notion de similarité. Pour cela, nous avons utilisé une approche fondée sur les réseaux de neurones siamois. Les résultats que nous avons obtenus montrent que le modèle de similarité appris au moyen de ces réseaux est capable de faire ressortir une différence significative entre les paires de voix doublant un même personnage et les autres paires de voix. Ces expériences montrent bien que cette différence ne peut être le fruit du hasard, il y a donc une information reliée à la dimension "personnage" sur laquelle s'appuie notre modèle.

Nous avons toutefois observé des variations dans nos 4 tests au niveau des scores moyens des groupes *target* et *nontarget*. La performance du modèle de similarité semble dépendre des personnages impliqués dans nos données. Il nous faut donc étudier plus en profondeur l'impact des différents personnages pour pouvoir être capable d'expliquer les possibles confusions du système. Par exemple en mettant en place un protocole sur plusieurs itérations, où un personnage différent est extrait du corpus à chaque fois. De manière générale, la mesure de similarité réalisée à l'aide des réseaux de neurones siamois apparaît plus pertinente qu'une approche inspirée de la reconnaissance du locuteur s'appuyant sur la PLDA (Gresse *et al.*, 2017). En effet, la PLDA a tendance à se focaliser sur le locuteur lui-même. À l'inverse, les réseaux de neurones siamois nous permettent d'apprendre un modèle sur la base de deux voix différentes. Nous considérons donc l'utilisation des réseaux de neurones siamois et l'apprentissage par paires de voix plus apte à la mesure d'une similarité. Le modèle tire à la fois parti des paires de voix doublant un même personnage et de celles doublant des personnages différents. Bien que nous observons une différence significative entre les scores *target* et *nontarget*, il nous est encore difficile d'expliquer cette différence. Nous consacrerons donc nos futurs travaux à l'approfondissement du travail présenté dans cet article. De plus, nous avons conscience que la méthode $i$-vecteur peut ne pas être la plus adéquate. En effet cette dernière est censée compenser une partie du bruit, mais cela peut nous amener à perdre de l'information utile pour caractériser certains aspects de la voix. La question de la représentation de l'information est donc une piste de recherche très intéressante que nous aborderons dans de futurs travaux.

# Remerciements

Les travaux présentés dans cet article sont financés par la Fondation de l'Université d'Avignon.

# Références


BAUMANN O. & BELIN P. (2010). Perceptual scaling of voice identity : common dimensions for different vowels and speakers. *Psychological Research PRPF*, **74**(1), 110.

BROMLEY J., GUYON I., LECUN Y., SÄCKINGER E. & SHAH R. (1994). Signature verification using a" siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, p. 737–744.

CHOPRA S., HADSELL R. & LECUN Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, p. 539–546 : IEEE.

DEHAK N., KENNY P. J., DEHAK R., DUMOUCHEL P. & OUELLET P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, **19**(4), 788–798.

GRESSE A., ROUVIER M., DUFOUR R., LABATUT V. & BONASTRE J.-F. (2017). Acoustic pairing of original and dubbed voices in the context of video game localization.

HADSELL R., CHOPRA S. & LECUN Y. (2006). Dimensionality reduction by learning an invariant mapping. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, p. 1735–1742 : IEEE.

KELLY F., ALEXANDER A., FORTH O., KENT S., LINDH J. & ÅKESSON J. (2016). Identifying perceptually similar voices with a speaker recognition system using auto-phonetic features. In *INTERSPEECH*, p. 1567–1568.

KOCH G., ZEMEL R. & SALAKHUTDINOV R. (2015). Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2.

LAVER J. (1980). The phonetic description of voice quality. *Cambridge Studies in Linguistics London*, **31**, 1–186.

LINDH J. & ERIKSSON A. (2010). Voice similarity-a comparison between judgements by human listeners and automatic voice comparison. In *Proceedings from FONETIK*, p. 63–69.

LOAKES D. (2006). *A forensic phonetic investigation into the speech patterns of identical and non-identical twins*. PhD thesis, University of Melbourne, School of Languages.

MCDOUGALL K. (2013). Assessing perceived voice similarity using multidimensional scaling for the construction of voice parades. *International Journal of Speech, Language & the Law*, **20**(2).

NOLAN F., FRENCH P., MCDOUGALL K., STEVENS L. & HUDSON T. (2011). The role of voice quality 'settings' in perceived voice similarity. *International Association for Forensic Phonetics and Acoustics, Vienna, Austria*.

OBIN N. & ROEBEL A. (2016). Similarity search of acted voices for automatic voice casting. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24**(9), 1642–1651.

OBIN N., ROEBEL A. & BACHMAN G. (2014). On automatic voice casting for expressive speech : Speaker recognition vs. speech classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, p. 950–954 : IEEE.

ROSE P. (1999). Differences and distinguishability in the acoustic characteristics of hello in voices of similar-sounding speakers. *Australian Review of Applied Linguistics*, **22**(1), 1–42.

ZHANG C. & TAN T. (2008). Voice disguise and automatic speaker recognition. *Forensic science international*, **175**(2), 118–122.

# Doubler les consonnes en chant baroque français : un cas de gémination expressive ?

Claire Pillot-Loiseau[1] Claudia Schweitzer[2] Christelle Dodane[3] Alice Romeo[1] Giuseppina Turco[1,4]

(1) Laboratoire de Phonétique et Phonologie (LPP) UMR 7018, Université Paris 3 Sorbonne Nouvelle, CNRS, 19 rue des Bernardins, 75005 Paris, France

(2) Histoire des théories linguistiques (HTL) UMR 7597, Université Paris 3 Sorbonne nouvelle, CNRS, et Université Paul Valéry Montpellier 3, Département musicologie, Route de Mende, 34199 Montpellier cedex 5, France

(3) PRAXILING UMR5267, Université Paul Valéry Montpellier 3, CNRS, Bâtiment Marc Bloch (BRED), Route de Mende, 34199 Montpellier cedex 5, France

(4) Laboratoire de Linguistique Formelle (LLF) UMR 7110, Université Paris-Diderot, CNRS, Case Postale 7031, 5, rue Thomas Mann, 75205 Paris Cedex 13, France

claire.pillot@sorbonne-nouvelle.fr, claudia.schweitzer2@gmail.com, christelle.dodane@univ-montp3.fr, alice.romeo@coursdiderot.com, gturco@linguist.univ-paris-diderot.fr

## RESUME

Quels sont les marqueurs acoustiques du *doublement de consonnes*, technique décrite au XVIII$^{ème}$ siècle en musique baroque vocale française, à des fins expressives ? Ces marqueurs s'assimilent-ils à la gémination ? Nous avons enregistré 5 chanteurs baroques produisant un air de Lully en parole et chant, sans et avec doublement consonantique. La durée des consonnes doublées rapportée à celle du mot de ces 50mn de productions a été mesurée, puis analysée statistiquement en fonction de la modalité (chanté vs. parlé), de la condition (avec doublement vs. sans doublement) et du type de consonne (/s,f,v/, /l,m/, /r/, et /p,t,k,b,d,g/). Nos résultats montrent une augmentation significative de la durée relative consonantique avec le doublement, surtout pour le chant, mais variant selon le chanteur et le type consonantique. En chant et parole, la voyelle précédente est plus courte quand elle est suivie d'une consonne doublée que quand elle ne l'est pas.

## ABSTRACT

**Doubling the consonants in French Baroque singing: Is it a case of expressive gemination?**

What are the acoustic cues of *consonant doubling*, a technique mainly described in the 18th century in French vocal baroque music, for expressive purposes? Do these markers assimilate to gemination? We recorded 5 Baroque singers producing a song by Lully in speech and song modalities, with and without consonants doubling. The analysis of the duration of the doubled consonants compared to that of the word of these 50mn productions was carried out, then analyzed statistically according to the modality (sung vs. spoken), the condition (with vs. without doubling) and the type of consonant ((/s,f,v/, /l,m/, /r/, and /p,t,k,b,d,g/). Our results show a significant increase in the relative consonantal duration with the doubling, especially for singing, but varying



10.21437/JEP.2018-3

according to the singer and the type of consonant. In singing and in speech, the preceding vowel is shorter when it is followed by a doubled consonant than when it is not.

MOTS-CLES : doublement consonantique, gémination, chant baroque, français, expressivité.
KEYWORDS: consonant doubling, gemination, Baroque vocal technique, French, expressivity

# 1 Introduction et état de l'art

Si les études actuelles foisonnent en matière de production des voyelles en chant lyrique monodique (entre autres : Sundberg 1990), moins de travaux concernent les consonnes chantées (entre autres : McCrea et Morris, 2005). Ces dernières études montrent que le *Voice Onset Time* (VOT) des consonnes /p, t, k, b, d, g/ produites par des chanteurs entraînés masculins anglophones est plus long que celui produit par des amateurs pour /p, t, k/, et est plus long en chant qu'en parole pour les mêmes trois consonnes. A la croisée des chemins entre l'histoire de la prononciation du français chanté, et nos techniques actuelles d'analyse acoustique, notre but est de conduire une investigation phonétique concernant la musique vocale baroque française, et en particulier au sujet de ce que les auteurs de l'époque appellent le « redoublement des consonnes » se traduisant par leur mise en valeur. A notre connaissance, aucune étude du genre n'a encore été conduite. Comment peut-on objectiver acoustiquement ce redoublement ? Est-il analogue à la gémination ?

## 1.1 La gémination

Se référant usuellement au dédoublement distinctif d'un phonème consonantique (Delattre 1971) facilement perceptible, la gémination est observée dans plusieurs langues (Turco *et al*. 2017). Cependant, l'allongement seul des consonnes ne suffit pas à décrire leur gémination (Ridouane 2010) : dans la plupart des langues, les voyelles précédant une consonne géminée sont plus courtes qu'avant une consonne simple. En outre, l'intensité de l'explosion des occlusives géminées, et la force de contact des articulateurs durant leur articulation (Ridouane 2007), sont plus importantes que pour des occlusives simples. Si la gémination permet une distinction phonologique, entre, par exemple, *fata* /ˈfata/ [ˈfaːta] « fée » et *fatta* /ˈfatta/ [ˈfat.ta] « faite » pour l'italien, l'allongement de la durée consonantique en français intervient dans l'accentuation d'insistance d'une part, et pour distinguer des énoncés comme *Elle a dit* /ɛl a di/ et *Elle l'a dit* /ɛl l‿a di/ d'autre part.

## 1.2 Le doublement de consonnes en chant baroque français

Prolongement et renforcement d'une consonne dans un mot, le doublement ou *redoublement* de consonnes dans le chant français est une mise en valeur d'un mot par une articulation renforcée de quelques consonnes, pratique dont parlent plusieurs sources de la deuxième moitié du XVIII$^{ème}$ siècle (Bérard 1755, Blanchet 1756, Lécuyer 1769, Raparlier 1772). Le chant étant considéré comme *« une déclamation plus embellie que la déclamation ordinaire »* (Bérard 1755 : 50), on comprend facilement l'importance que les auteurs accordent à la bonne prononciation du chanteur, dans laquelle les consonnes jouent un rôle essentiel. Pour visualiser la consonne redoublée, Bérard (1755), Blanchet (1756) et Raparlier (1772) utilisent deux fois la même lettre : 1) dans le mot même, 2) au-dessus de la première lettre (figure 1). Les auteurs indiquent qu'en cas de redoublement consonantique, le son de la consonne est prolongé, soutenu dans le temps. Physiquement, les organes continuent leur mouvement articulatoire ce qui entraîne l'expression du *redoublement*. Son degré (en intensité et longueur) dépend du contenu du texte : il est plus fort pour

les émotions violentes que pour les sentiments doux et tendres. Cette prolongation peut aussi être réalisée par une préparation consciente de la consonne en question qui, à ce dessein, est « préparée » ou « retenue »[1] un petit moment. Ces deux actions s'accompagnent d'un contrôle du processus de l'articulation et de l'expiration selon le caractère voulu.

Ces auteurs soulignent cependant l'importance de l'emploi de cette technique à des fins expressives *« on doit doubler les lettres dans tous les endroits marqués au coin de la passion »* (Raparlier 1772). En témoigne le commentaire initiant la partition de chaque air illustré dans le traité de Bérard (exemple figure 1) : parlant des *sons à caractère* pour parler des *sons violens*, *sons entrecoupés*, *sons majestueux*, *sons légers* ou *sons tendres et délicats,* l'auteur, pour les *sons violens* incarnés par l'expression de l'agitation du personnage Atys (voir figure 2 pour les paroles de ce récitatif), recommande au chanteur de « faire sortir avec une extrême rapidité l'air intérieur, prononcer d'une manière dure et obscure, et doubler assez fortement les lettres » (figure 1). En revanche, pour un son léger, *« il faut chasser l'air interieur en petit volume, expirer peu de temps pour les divers sons, et préparer tres foiblement les lettres »* (Bérard 1755 : Annexe : 21).

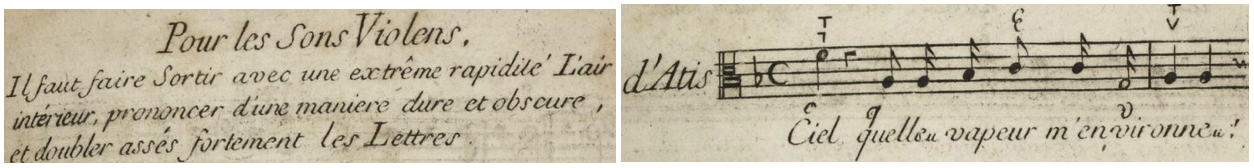



FIGURE 1: Recommandations (gauche) et indication (droite) des consonnes doublées par Bérard (1755)[2] dans le début du récit « Ciel, quelle vapeur m'environne » d'*Atys* (acte V, scène 3) de Lully (1676) : les consonnes doublées (soulignées ci-dessus) se situent entre le texte et la portée musicale.

Notre but est de savoir si, dans notre corpus, la durée est un paramètre acoustique pertinent et analogue à ces modifications rencontrées dans la gémination. Pour décrire les consonnes doublées, nous présentons l'analyse acoustique du récitatif de la figure 1, produit en voix parlée et chantée sans et avec redoublement des consonnes, récitatif choisi 1) pour être souvent cité par Bérard et Blanchet (1755, 1756) afin d'expliquer la technique du doublement consonantique ; 2) parce qu'il exprime de violentes passions d'ordinaire associées à un grand nombre de consonnes doublées.

# 2 Méthode

Six chanteurs français ont été enregistrés[3]. Nous présentons ici l'analyse de cinq d'entre eux (âge moyen des 3 femmes (soprani) et des deux hommes (un contre-ténor et un baryon) : 41,4 ans ; ET : 14,5 ; de 27 à 63 ans). Trois sont professionnels, un est semi-professionnel et une est enseignante à l'université. Quatre enseignent le chant pratiqué depuis en moyenne 13 ans (ET: 5,8). Ces sujets effectuent de 4 à 25 concerts par an en temps que soliste (ensemble vocal pour trois d'entre eux). Tous pratiquent le répertoire baroque pour lequel ils ont reçu une formation spécifique (cours particuliers, masterclass, stages, Centre de Musique Baroque de Versailles, université) ; deux d'entre eux ont particulièrement étudié le doublement de consonnes. Soulignons que le recrutement bénévole de personnes possédant ces compétences spécifiques ne fut pas aisé, d'où le petit nombre

[1] « Les personnes, émues par quelque passion doublent, ou (ce qui est le même) préparent ou retiennent ordinairement les consonnes dans l'Articulation » (Bérard 1755:93; Blanchet 1756:53).

[2] Traité librement accessible en ligne au lien suivant : http://gallica.bnf.fr/ark:/12148/btv1b8623287n

[3] Le détail des questions posées aux sujets et leurs réponses peuvent être visualisés au lien suivant : https://docs.google.com/spreadsheets/d/1feVkXKg4dTPdPCL3og1Ptch0wPa3nUZY_tKMvA8Zdnc/edit?usp=sharing

de sujets ayant participé à cette étude. Les sujets, ayant tous signé un formulaire de consentement, ont été enregistrés debout ou assis en chambre sourde avec un microphone électrostatique serre-tête (AKG C520L positionné à 5cm de la bouche du sujet), et avec le logiciel ProTools (fréquence d'échantillonnage : 44100Hz).

Deux airs parlés et chantés ont été enregistrés deux fois sans (normalement) puis avec doublement consonantique (en doublant cette fois-ci les consonnes au-dessus du texte). Nous présentons l'analyse du premier air (figures 1 et 2) contenant 23 fricatives, 12 /r/ (prononcée ici comme vibrante), 14 occlusives, 5 /l/ et un /m/ (figure 2). Des extraits musicaux plutôt que des phrases de laboratoire aux contextes contrôlés ont été choisis, afin de garantir une expressivité authentique dans les productions des chanteurs. Chaque sujet pouvait chanter l'air dans la tonalité qu'il voulait, mais devait garder le même ton dans les contextes sans et avec doublement. Pour cette dernière condition, il leur était juste demander de « doubler » les consonnes sans autre indication. Pour assurer une homogénéité dans les productions obtenues, il a été demandé aux sujets de parler et chanter l'air dans la prononciation actuelle, en doublant toutefois les consonnes dans le contexte « doublé » (sans indiquer de quelle façon il fallait doubler les consonnes). Le gain a été maintenu constant pour un même contexte (parlé ou chanté, première ou deuxième répétition) sans et avec doublement. La totalité des productions dure en moyenne 10mn29 par chanteur (écart-type : 43s) mais les temps d'échanges ont abouti à un enregistrement total d'en moyenne une heure par sujet. 52mn26s de productions parlées et chantées ont ensuite été segmentées phonétiquement à l'aide de la plateforme *Munich AUtomatic Segmentation* (MAUS : Kisler et al. 2017), puis vérifiés manuellement (une première fois par quatre des auteurs, puis une deuxième fois par un seul auteur, pour assurer une parfaite homogénéité de la segmentation), selon les critères établis par Turk *et al.* (2006). En cas de liaisons (ex. entre *tout* et *à coup*), le découpage des mots s'est effectué selon les règles phonotactiques du français (tout-ta-coup).

C q v
Ciel ! Quelle vapeur m'environne !
s s b r r r l t r rd
Tous mes sens sont troublez, je frémis, je frissonne ; je tremble, et tout à coup, une infernale ardeur
f s d c
Vient enflamer mon sang, et dévorer mon coeur.
i c s r
Dieux ! Que vois-je ? Le ciel s'arme contre la terre,
r l t
Quel désordre ! Quel bruit ! Quel éclat de tonnerre !
r f s s t
Quels abimes profonds sous mes pas sont ouverts !
t s s f
Que de phantosmes vains sont sortis des Enfers !
s g f l q d rb c r s p c r
Sangaride, ah fuyez la mort que vous prépare, une divinité barbare, c'est votre seul péril qui cause ma terreur.
m t f l
Quel monstre vient à nous ! Quelle fureur le guide !
s r l S
Ah respecte, cruel, l'aimable Sangaride.

Quels hurlements affreux !
f s c
Il faut combattre, amour, seconde mon courage.

| | | | |
|---|---|---|---|
| Fricatives (n=23) | /s/ 16 | Occlusives (n=14) | /p/ 1 |
| | /f/ 6 | | /b/ 2 |
| | /v/ 1 | | /t/ 5 |
| /r/ vibrante (n=12) | /r/ 12 | | /d/ 1 |
| /l, m/ (n=6) | /l/ 5 | | /k/ 4 |
| | /m/ 1 | | /j/ 1 |

FIGURE 2: paroles du récitatif « Ciel, quelle vapeur m'environne » de *Atys* de Lully (1676) et comptage des consonnes doublées par type : les consonnes doublées apparaissent sur le texte[4].

[4] La consonne et la semi-consonne entourées ont été retirées de l'analyse en raison d'une pause avant l'occlusive sourde, interdisant d'en calculer la durée. La semi-consonne n'a pas été analysée car elle se trouve en dehors des types consonantiques ici étudiés.

Les durées relatives des consonnes doublées et de leur voyelle précédente (durée / durée du mot), ont ensuite été calculées avec un script Praat (Boersma et Weenink 2017). Nous calculons ces durées relatives car nous comparons le chant et la parole, avec des vitesses d'élocution potentiellement différentes (Picket *et al*. 1999). La durée de la voyelle précédente a été choisie car son raccourcissement (pour l'italien par exemple) ou son allongement (comme en Japonais) durant la gémination, a été interprété comme la conséquence d'une articulation plus tendue requise par la géminée suivante, faisant de cette mesure un autre indice temporel de la gémination (Ridouane 2010). A l'exception de la consonne entourée sur la figure 2, toutes les occlusives sourdes cible ont été produites sans pauses par nos sujets. Un modèle linéaire mixte (bibliothèque 'lme4', Bates *et al.*, 2014) a permis d'étudier les relations entre la durée consonantique relative, celle de la voyelle précédente et les facteurs fixes *modalité* (chanté/parlé), *condition* (sans/avec doublement), et *type consonantique* (/s,f,v//l,m//r/ /p,t,k,b,d,g/).

# 3 Résultats

## 3.1 Durée relative des consonnes doublées

Les facteurs modalité ($\chi2(1)=12.20$, $p<0,0001$), condition ($\chi2(1)=12.43$ $p<0,0001$) et type de consonne ($\chi2(3)= 12.26$, $p<0,001$) ont un effet significatif sur la durée consonantique relative.

Les durées consonantiques relatives sont plus importantes en parole qu'en chant (*modalité* : $\beta_{parlé}=0,11$, $SE=0,01$, $t=5,94$, $p<0,0001$), sauf pour /l,m/ chez le chanteur S4. En outre, les consonnes sont plus longues *avec doublement* que *sans* celui-ci (*condition* : $\beta_{sans\ doubl.}=-0,04$, $SE=0,006$, $t=-7,12$, $p< 0,0001$ excepté /r/ en parole chez S3, /l,m/ chez S4 dans ce même contexte, ainsi que chez S5), avec cependant une importante variabilité, surtout en chant (figure 3[5]). De plus, l'augmentation de durée consonantique relative *avec doublement* est, d'après des comparaisons appariées, davantage significative en chant qu'en parole, par rapport à la condition *sans doublement*.

Concernant le *type consonantique*, les durées relatives de /r/ - le plus doublé pour 3 sujets sur 5 (sauf S2 et S3) - sont les moins variables. /l, m/ augmentent le moins leur durée consonantique relative avec le doublement en chant. Le doublement des fricatives en voix parlée n'entraîne jamais d'augmentation significative de cette durée, contrairement au chant, surtout pour S2, S3 et S5. La durée relative des occlusives augmente également significativement de la condition *non doublé* à *doublé*, surtout chez S3, S4 et S5. Indépendamment de la condition, /f,s,v/ et /l,m/ sont significativement plus allongées que /p,t,k,d,g,b/ ($\beta_{/p,t.k,b,d,g/}=-0,09$, $SE=0,01$, $t=-7,52$, $p<0,0001$; $\beta_{/p,t.k,b,d,g/}=0,04$, $SE=0,02$, $t=2,20$, $p< 0,05$) et /r/ ($\beta_{/r/}=0,07$, $SE=0,01$, $t= 4,14$, $p< 0,0001$).

Nous notons également que les résultats sont variables selon les sujets : S4 (seul chanteur formé au Centre de Musique Baroque de Versailles) est celui qui double le plus en parole et chant, ainsi que S3 et S5 en chant, autres sujets ayant une formation baroque plus avancée.

La figure 3 montre une importante variabilité des durées consonantiques relatives obtenues, variabilité notamment due aux différents contextes phonétiques dans lesquels se trouvent les consonnes que nous étudions (segment précédent et suivant, position prosodique, mais aussi fréquence fondamentale, surtout en chant). Les occurrences correspondant aux augmentations de durée relative consonantique les plus importantes des conditions *sans doublement* à *avec*

[5] Par souci de clarté, la modalité parlée n'y est pas représentée. Le lecteur pourra télécharger cette information au lien suivant : https://drive.google.com/open?id=1sucR--140wqtdEAnqP__OGUUTQ5Fjzwx

*doublement* pour tous les sujets, sont : les fricatives des mots « seconde, respecte, enflammer, Sangaride », les /l/ de « le, la », le /r/ de « bruit » et les occlusives de « dévorer, tonnerre » en modalité parlée. Concernant la modalité chantée, le /s/ de « seul », les /l/ de « le, la », les /r/ de « bruit, barbare, tremble, infernale, votre » ainsi que les occlusives de « que, tout » ont été les plus allongées des conditions *sans doublement* à *avec doublement*.

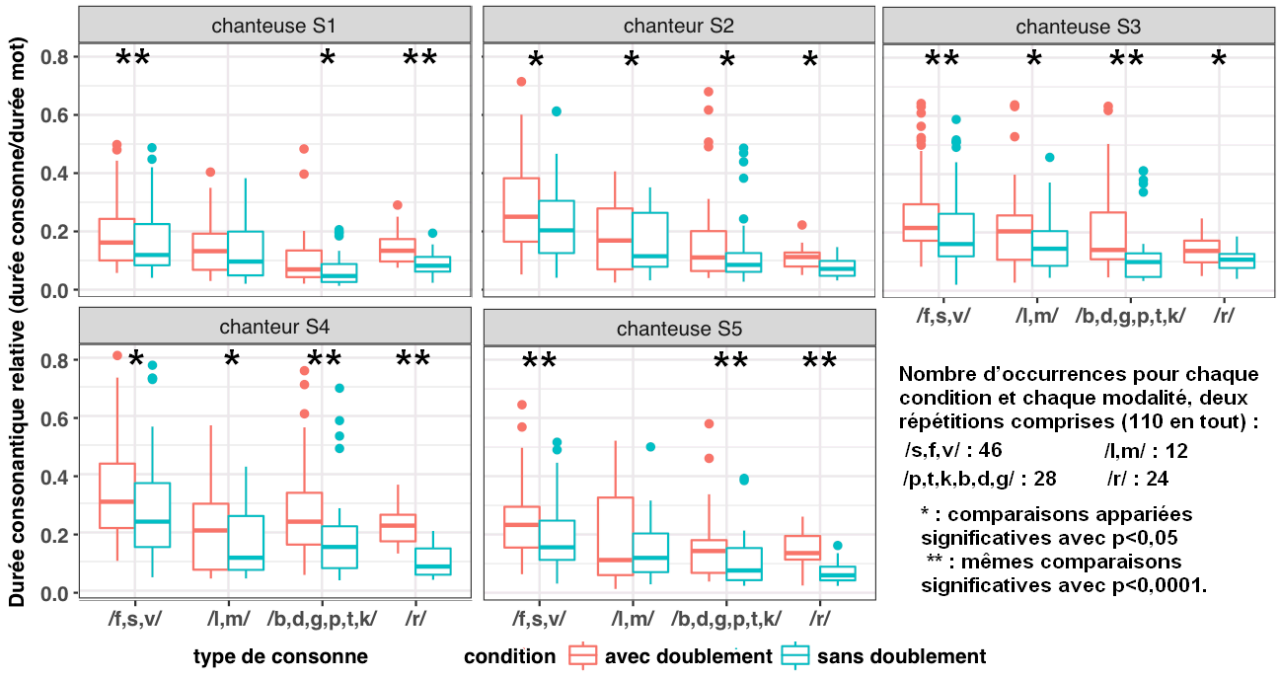



FIGURE 3: Durées consonantiques relatives (s) en modalité chantée en fonction du sujet, de la condition (sans/avec doublement), et du type consonantique (/s,f,v/, /l,m/, /p,t,k,b,d,g/, et /r/).

## 3.2 Durée de la voyelle précédant les consonnes doublées

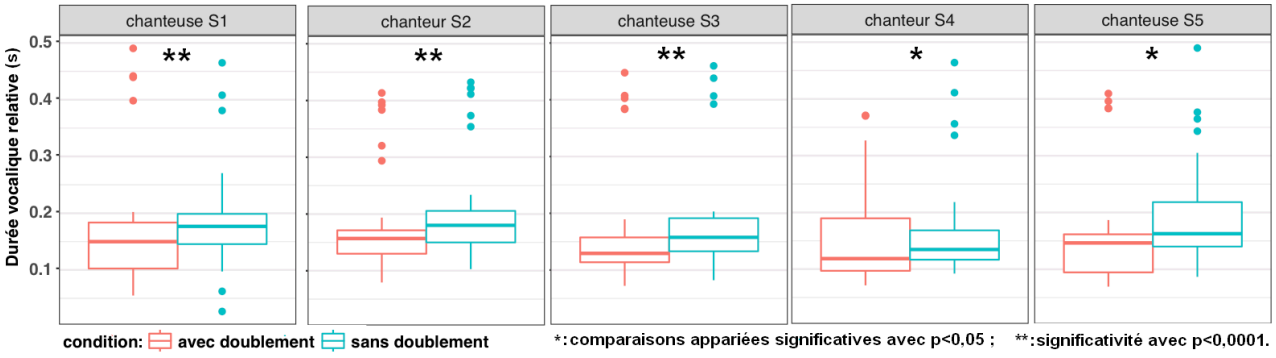



FIGURE 4 : Durées relatives de la voyelle précédant chaque consonne doublée (s) en modalité chantée, en fonction du sujet et de la condition (sans/avec doublement)

Le facteur condition (sans doublement/avec doublement) a un effet significatif sur la durée vocalique relative ($\chi2(1)=8,08$, $p<0,001$), alors qu'aucun effet de la modalité (parlé/chanté) n'a été obtenu ($p=2$) : excepté en modalité parlée chez S5, toutes les voyelles précédant une consonne doublée sont significativement plus courtes en condition *avec doublement* que sans celui-ci ($\beta_{sans\ doubl}=0,03$, $SE=0,01$, $t=2,79$, $p<0,001$), que ce soit en modalité parlée ou chantée (figure 4[5]).

# 4 Discussion

Nos résultats montrent que, comme attendu, le doublement de consonnes (*condition*) entraîne pour tous une augmentation de la durée consonantique relative et une diminution de celle de la voyelle précédant les consonnes doublées. Ce phénomène se produit tant en parole qu'en chant : notons en effet que Bérard (1755) et Blanchet (1756) donnent aussi des exemples de tragédies, donc de textes parlés, nécessitant également un doublement consonantique. Ces paramètres temporels semblent donc, comme la gémination, également robustes pour produire une consonne doublée vs. non doublée chez nos sujets. Ridouane (2003) synthétise les paramètres acoustiques permettant de décrire la gémination pour les occlusives de 27 langues : la durée de l'occlusion, en général trois fois plus importante pour les occlusives géminées que pour leurs équivalentes simples, est le paramètre acoustique caractérisant le plus les occlusives géminées pour tous ces parlers (également perceptivement), quelle que soit la consonne occlusive et sa position dans le mot. En outre, la durée de la voyelle précédant la consonne géminée est parfois moins longue devant cette consonne géminée que devant son équivalente simple. Pour certaines langues seulement, la durée de l'explosion, mais aussi son intensité, et la durée du *VOT*, peuvent être majorées et plus importantes pour les consonnes occlusives géminées. Enfin, il a été rapporté que le premier formant des voyelles environnantes était plus élevé (voyelles plus ouvertes) en présence d'occlusives géminées. Pourtant, au cours de nos échanges avec nos sujets où les même questions 18 et 19 ont été posées[3], nos locuteurs ont librement exprimé la façon dont ils pensaient doubler les consonnes, et l'avantage que cette technique représentait pour eux : selon eux, ils produisent les occlusives doublées de manière plus percussive, directive et tonique, mais aussi anticipée et accentuée par l'intensité. Pour notre population, le doublement aide à mieux chanter dans le corps par l'énergie requise, et permet de projeter la voyelle suivante et la voix en général. Il peut également créer un rythme dans ce récitatif. Nos sujets n'ont pas mentionné un allongement consonantique et une réduction de la voyelle précédant la consonne doublée comme phénomène à la base du doublement qu'ils ont produit, mais ils ont évoqué d'autres critères que les paramètres temporels que nous avons choisis dans nos mesures, comme l'intensité (Kawahara 2015).

Les durées consonantiques relatives en parole sont plus importantes qu'en chant (*modalité*), en condition *doublé* ou *non doublé* : les chanteurs ne sont pas contraints par un rythme donné en parole contrairement au chant. Bien que ce chant choisi soit un récitatif donc plus proche du parlé qu'un air, son rythme chanté est plus lent qu'en parlé (ex: version chantée *doublée* de S4 : débit de parole : 137 syllabes/minute ; version parlée *doublée* de S4 : 150 syllabes/minute) : le débit de parole étant plus rapide en parole qu'en chant, les durées relatives consonantiques sont probablement majorées dans cette modalité. Nous pourrons à l'avenir affiner nos mesures rythmiques de ces modalités pour confirmer cette interprétation de nos résultats.

Concernant le *type consonantique*, la durée relative de toutes les consonnes augmente de la condition *non doublé* à *doublé*, avec un phénomène moins marqué pour /l,m/. Peu de travaux sur la gémination, dont aucun en chant, étudient d'autres consonnes que les occlusives. Payne (2005) obtient des durées de /f/ géminés les plus importantes en parole italienne, comme nos productions françaises de /s,f,v/ parlées, et, à un moindre degré, chantées. Ses occlusives sourdes géminées sont plus longues que les voisées. Ses consonnes /l,m/ font partie des géminées moins longues.

Il reste que la fonction de ce doublement consonantique est bien d'ordre expressive : dans ce récitatif où le personnage Atys exprime des passions violentes, les plus importants doublements de /r/ dans *tremble, bruit, barbare, infernale*, celui des fricatives dans *enflammer*, figuralisent idéalement les passions exprimées par ces mots, accompagnées de l'emphase des mots

grammaticaux *le, la, que, tout*, qu'il est possible d'assimiler à une « gémination expressive » comme observée dans certains parlers allemands (Sturm 2016), mais aussi aux phénomènes d'allongement liés à l'accentuation initiale en français, tout particulièrement en initiale de mot comme le sont la plupart des occurrences de consonnes doublées dans ce récitatif (Jun & Fougeron 2002).

# Conclusion

Nos résultats montrent que la modification des paramètres temporels chez nos sujets en parole et chant baroque français, s'assimile à la gémination observée dans la plupart des langues (allongement de la durée consonantique et raccourcissement de la durée de la voyelle précédant une consonne géminé). Il conviendrait à l'avenir, en complément de ces paramètres temporels, 1° d'affiner leur analyse en fonction des contextes phonétiques et prosodiques des consonne cibles et de la nature de la voyelle précédente ; 2° d'analyser l'intensité relative des consonnes doublées et non doublées ; 3° de mesurer les formants de /l,m/, ainsi que le premier formant des voyelles environnant les consonnes cibles (Kawahara 2015) car nos sujets affirment percevoir une modification de la qualité vocalique avec le doublement, même si les formants vocaliques ne se modifient pas dans tous les phénomènes de gémination (Esposito & Di Benedetto 1999).

# Remerciements

Nous remercions les sujets d'avoir accepté bénévolement de se faire enregistrer. En outre, ce travail a bénéficié d'une aide du LabEx EFL (ANR-10- LABX-0083).

# Références

BATES, D., MAECHLER, M., BOLKER, B., & WALKER, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R package version*, *1*(7), 1-23.

BERARD, J-A. (1755*). L'Art du chant*. Paris : Dessaint & Saillant, Prault et Lambert.

BOERSMA, P., WEENINK, D. (2017). Praat: doing phonetics by computer Version 6.0.16, retrieved 3 March 2017 from http://www.praat.org/

BLANCHET, J. (1756). *L'Art ou les Principes philosophiques du chant.* Paris : Lottin, Lambert et Duchesne.

DELATTRE, P. (1971). Consonant gemination in four languages: an acoustic, perceptual and radiographic study, part I. *IRAL-International Review of Applied Linguistics in Language Teaching*, *9*(1), 31-52.

ESPOSITO, A., DI BENEDETTO, M. G. (1999). Acoustical and perceptual study of gemination in Italian stops. *The Journal of the Acoustical Society of America*, *106*(4), 2051-2062.

JUN, S. A., FOUGERON, C. (2002). Realizations of accentual phrase in French intonation. *Probus*, *14*(1), 147-172.

KAWAHARA, S. (2015). The phonetics of *sokuon*, or geminate obstruents, in Haruo Kubozono (Eds.) Handbook of Japanese Phonetics and Phonology. Berlin : de Gruyter, 43-77.

KISLER, T., REICHEL U. D., SCHIEL, F. (2017). Multilingual processing of speech via web services, Computer Speech & Language, Volume 45, September 2017, pages 326–347.

MCCREA, C., MORRIS, R. (2005). Comparison of voice onset time for trained male singers and male nonsingers during speaking and singing. *Journal of voice,* 19(3) : 420-430.

PAYNE, E.M. (2005). Phonetic variation in Italian consonant gemination. *Journal of the International Phonetic Association*, *35*(2), 153-181.

PICKETT, E.R., BLUMSTEIN, S.E., BURTON, M.W. (1999). Effects of speaking rate on the singleton/geminate consonant contrast in Italian. *Phonetica*, *56*(3-4), 135-157.

R DEVELOPMENT CORE & TEAM. R: A language and environment for statistical computing (Version 2.15.0). Austria: The R Foundation for Statistical Computing. Retrieved from http://www.r-project.org.

RAPARLIER, A-A. (1772). *Principes de la musique, les agréments du chant et un essai sur la prononciation, l'articulation et la prosodie de la langue française.* Lille : Lalau.

RIDOUANE, R. (2007). Gemination in Tashlhiyt Berber: an acoustic and articulatory study. *International Phonetic Association*. *Journal of the International Phonetic Association*, *37*(2), 119.

RIDOUANE, R. (2010). Geminates at the junction of phonetics and phonology. *Laboratory phonology*, *10*, 61-90.

STURM, L. (2016). Expressiveness and variation: the etymology of German Kladder 'Dirt, mud'. *Studia Linguistica Universitatis Iagellonicae Cracoviensis*, *133*(2), 109-114.

SUNDBERG J. (2009). Articulatory Configuration and Pitch in a Classically Trained Soprano Singer. *Journal of Voice*, 23(5), 546-551.

TURCO, G., SHOUL, K., RIDOUANE, R. (2017). How are four-level length distinctions produced? Evidence from Moroccan Arabic. *Proc*. *Interspeech 2017*, 215-218.

TURK, A., NAKAI, S., & SUGAHARA, M. (2006). Acoustic segment durations in prosodic research: A practical guide. *Methods in empirical prosody research*, *3*, 1-28.

# Comparaison des voix dans le cadre judiciaire : influence du contenu phonétique

Moez Ajili[1] Jean-François Bonastre[1] Waad Ben Kheder[1] Solange Rossato[2] Juliette Kahn[3]
(1) Univ. Avignon, LIA, F-84000 Avignon France
(2) Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France
(3) LNE, F-78000 Trappes France
(1) prenom.nom@univ-avignon.fr (2) prenom.nom@univ-grenoble.fr, (3) prenom.nom@lne.fr

RÉSUMÉ
En comparaison de voix dans le domaine criminalistique, l'approche Bayésienne est devenue le nouveau "golden standard". Dans cette approche, l'expert exprime ses résultats par un unique nombre, le rapport de vraisemblance (LR). Cet article s'intéresse à l'influence du contenu phonétique sur la fiabilité du LR. Nous nous intéressons particulièrement à la quantité d'information spécifique au locuteur que portent les différents sons de la parole. Cette étude met en évidence des différences importantes entre les phonèmes et, surtout, la forte influence de la variabilité intra-locuteur.

ABSTRACT
**phonetic content impact on forensic voice comparison.**

Forensic Voice Comparison (FVC) is increasingly using the *likelihood ratio* ($LR$).This article focuses on the impact of phonemic content on FVC performance and variability. The results demonstrate the importance of the phonemic content and highlight interesting differences between inter-speakers effects and intra-speaker's ones.

MOTS-CLÉS : Reconnaissance du locuteur, comparaison de voix, crimnalistique, fiabilité, contenu phonémique..

KEYWORDS: Forensic voice comparison, phonemic category, reliability..

# 1 Introduction

Dans les procédures judiciaires la comparaison de voix -ou "Forensic Voice Comparison (FVC)"- est de plus en plus fréquemment employée. L'approche Bayésienne est devenue le nouveau "golden standard" en sciences criminalistiques (Providers, 2009; Champod & Meuwly, 2000; Aitken & Taroni, 2004). Dans cette approche, l'expert exprime le résultat de son analyse sous la forme d'un unique nombre, le rapport de vraisemblance (LR) :

$$LR = \frac{\mathrm{p(E \mid H_p)}}{\mathrm{p(E \mid H_d)}} \quad (1)$$

où E est la trace, $H_p$ est l'hypothèse de culpabilité (même origine), et $H_d$ est l'hypothèse d'innocence (différentes origines).


10.21437/JEP.2018-4

Ce rapport ne favorise pas seulement une des hypothèses (“culpabilité” ou “innocence”) mais il fournit également le poids de ce support. Cet article poursuit les travaux présentés dans (Ajili *et al.*, 2016b,c). Nous nous visons à hiérarchiser les catégories phonétiques des sons de parole selon la quantité d'information spécifique au locuteur qu'elles contiennent. Le système automatique de reconnaissance du locuteur est utilisé comme outil de mesure dans cette étude.

Cet article présente en section 2 une vue de la littérature consacrée à l'influence du contenu phonétique sur la caractérisation du locuteur. Le protocole expérimental est présenté dans la section 3. La section 4 présente les expériences en découlant et les résultats associés. Un focus est proposé sur l'influence de la bande passante. Enfin, la section présente des conclusions et perspectives.

# 2 Contenu phonétique et caractérisation du locuteur

L'étude de l'information spécifique du locuteur portées par les phonèmes individuels ou encore des classes de phonèmes a fait l'objet de différents travaux comme (Wolf, 1972; Sambur, 1975; Eatock & Mason, 1994; Hofker, 1977; Kashyap, 1976; Amino *et al.*, 2006, 2012; Antal & Toderean, 2006). Les voyelles orales et les nasales apparaissent en tête en termes de discrimination entre les locuteurs. /s/, /t/ et /b/ sont souvent évalués comme moins porteurs d'information spécifique que les voyelles et les nasales. (Magrin-Chagnolleau *et al.*, 1995) utilise déjà un système automatique de reconnaissance du locuteur pour évaluer le pouvoir discriminant des différents phonèmes. Les auteurs suggèrent que les glissements et les liquides ensemble, les voyelles - et plus particulièrement les voyelles nasales - et les consonnes nasales contiennent plus d'informations spécifiques aux locuteurs qu'un enregistrement vocal phonétiquement équilibré. (Besacier *et al.*, 2000; Gallardo *et al.*, 2014) s'appuient également sur un système automatique. Ils montrent que certaines sous-bandes de fréquence sont plus pertinentes pour caractériser les locuteurs que d'autres, soulignant ainsi l'importance de la bande passante.

# 3 Protocole expérimental

Cette section est dédiée au protocole expérimental original utilisé pour cette étude, qui s'appuie sur le corpus FABIOLE. Ce corpus est distribué publiquement et a été créé dans le cadre du projet FABIOLE ANR-12-BS03-0011. FABIOLE (Ajili *et al.*, 2016a) présente des caractéristiques dédiées à l'étude de la variabilité intra-locuteur. Les extraits de parole proviennent d'émissions de radio ou de télévision françaises, avec une bonne qualité sonore et présentent une durée minimale de 30 secondes de parole obtenue par concaténation de segments issus de la même émission. FABIOLE est composée d'enregistrements venant de différents types de locuteurs, incluant des journalistes, des présentateurs, des politiciens, etc. Les contenus de FABIOLE sont proches de ceux des bases REPERE (Giraudel *et al.*, 2012), ESTER 1, ESTER 2 (Galliano *et al.*, 2005) et ETAPE (Gravier *et al.*, 2012). Cette caractéristique permet d'utiliser ces bases pour l'entraînement des modèles. FABIOLE contient des enregistrements prononcés par 130 locuteurs masculins, français natifs, répartis en deux ensembles : $T$ : 30 locuteurs cibles associés chacun à 100 extraits de parole ; $I$ : 100 locuteurs imposteurs associés chacun à un seul extrait de parole. Les données de FABIOLE ont été automatiquement transcrites pour réaliser un étiquetage phonétique en utilisant le système Speeral (Linares *et al.*, 2007).

Dans cet article, seul l'ensemble $T$ est utilisé.Pour chaque locuteur de $T$, 294950 paires d'enregistrements sont constituées. $4950$ de ces paires sont des comparaisons cibles et $290k$ des comparaisons imposteurs. Les paires cibles sont obtenues en utilisant toutes les combinaisons des 100 enregis-

trements disponibles pour chaque locuteur alors que les comparaisons non-cibles apparient chaque enregistrement du locuteur cible en question (100 sont disponibles) avec chacun des enregistrements des 29 locuteurs restants, formant par conséquent $(100 \times 100 \times 29 = 290k)$ comparaisons imposteurs.

Le système de reconnaissance du locuteur utilisé est LIA_SpkDet (Matrouf *et al.*, 2007) développé avec ALIZE (Bonastre *et al.*, 2005, 2008; Larcher *et al.*, 2013), qui met en œuvre une approche I-vector (Dehak *et al.*, 2011). Les paramètres acoustiques sont composés de 19 LFCC, de leur dérivées et de 11 dérivées secondes. Une bande passante réduite à la bande téléphonique (300-3400 Hz) est utilisée pour rester proche des conditions classiques en criminalistique. Cependant, à des fins de comparaison, une bande large tirant pleinement partie de la haute qualité des enregistrements d'origine est utilisée pour une des expériences.

Un *Universal Background Model* (*UBM*) de 512 composantes a été entraîné sur Ester 1&2, REPERE et ETAPE, en utilisant des locuteurs hommes qui n'apparaissent pas dans FABIOLE. La "total variability matrix" nécessaire à l'extraction des *i-vectors* a été apprise sur les mêmes données et, enfin, un modèle PLDA est utilisé pour le scoring (Prince & Elder, 2007).

Le $C_{llr}$ et le minimum de $C_{llr}$, dénoté $C_{llr}^{min}$, sont utilisés comme mesures de performance (Morrison, 2009; Brümmer & du Preez, 2006; Castro, 2007; Gonzalez-Rodriguez & Ramos, 2007). Le $\mathrm{C}_{llr}$ est défini par :

$$\mathrm{C_{llr}} = \underbrace{\frac{1}{2\mathrm{N_{tar}}} \sum_{\mathrm{LR} \in \chi_{\mathrm{tar}}} \log_2 \left(1 + \frac{1}{\mathrm{LR}}\right)}_{\mathrm{C_{llr}^{TAR}}} + \underbrace{\frac{1}{2\mathrm{N_{non}}} \sum_{\mathrm{LR} \in \chi_{\mathrm{non}}} \log_2 \left(1 + \mathrm{LR}\right)}_{\mathrm{C_{llr}^{NON}}} \tag{2}$$

$C_{llr}$ a la signification d'un coût ou d'une perte d'information (plus petit est le $C_{llr}$, meilleure est la performance) et peut être décomposé en deux parties additives :

— $\mathrm{C_{llr}^{TAR}}$, qui correspond à la perte moyenne relative aux paires target (le même locuteur a prononcé les deux enregistrements).

— $\mathrm{C_{llr}^{NON}}$, qui correspond à la perte moyenne relative aux paires non-target (les deux enregistrements proviennent de deux locuteurs différents).

Les scores issus d'un système automatique de reconnaissance de locuteurs doivent être "calibrés" pour devenir des LR. La transformation affine (Brümmer *et al.*, 2007) est employée pour cela, estimée en utilisant toutes les paires disponibles.

Nous avons choisi de travailler au niveau de classes de phonèmes en utilisant la classification suivante : Oral Vowels (OV) (/i/, /y/, /u/, /e/, /ø/, /o/,/ɛ/, /œ/,/ɔ/, /a/) ; Nasal vowels (NV) (/ɑ̃/, /ɔ̃/, /œ̃/, /ɛ̃/.) ; Nasal consonants (NC) ( /m/, /n/) ; Plosives (P) (/p/, /t/, /k/, /b/, /d/, /g/) ; Fricatives (F) (/f/, /s/, /ʃ/, /v/, /z/, /ʒ/) et Liquides (L) (qui comprend /l/, /ʁ/).

Pour déterminer l'influence d'une classe phonétique spécifique, nous utilisons une stratégie de "knock-out" : la part de signal correspondant à la catégorie étudiée est retirée des enregistrements et la perte de performance indique alors l'influence de celle-ci. Par conséquence, nous réalisons tout un jeu d'expériences dans lesquelles le matériel de parole correspondant à une classe phonétique spécifique est retiré des deux enregistrements composant une paire de comparaison de voix. La condition expérimentale correspondante est dénommée "**Specific**". La quantité de données correspondant à une catégorie spécifique est fortement variable d'une catégorie à l'autre mais également d'un enregistrement à l'autre (par exemple, dans nos expériences, les consonnes nasales représentent

6% du signal de parole alors que les voyelles orales pèsent pour 36%). Pour pallier ce biais, nous créons une condition de contrôle dénommée "**Random**", où la même quantité de signal est supprimée aléatoirement des enregistrements. Plus précisément, pour chaque enregistrement, quand un certain pourcentage de trames de parole est supprimé pour la condition '**Specific**", le même pourcentage est sélectionné aléatoirement et supprimé pour la condition "**Random**". Par précaution, ce processus est répété 20 fois, créant 20 fois plus de paires dans la condition "**Random**" que pour la condition "**Specific**".

L'impact d'une classe phonétique donnée est estimé relativement par $C_{\text{llr}}^{R}$ :

$$C_{\text{llr}}^{R} = \frac{\text{Cllr}^{\text{random}} - \text{Cllr}^{\text{specific}}}{\text{Cllr}^{\text{random}}} \times 100\% \tag{3}$$

Une valeur positive de $C_{llr}^{R}$ indique que la classe phonétique étudiée apporte moins d'information spécifique du locuteur que la condition de controle, une valeur négative montre au contraire que cette classe phonétique apporte plus d'information spécifique du locuteur que le contenu moyen.

# 4 Expériences et résultats

Nous avons tout d'abord calculé pour référence la performance sur l'ensemble des locuteurs et sur les contenus complets. Le $\text{C}_{\text{llr}}$ global correspondant est de 0.12631 bits (et l'EER de 2.88%).

TABLE 1 – $\text{C}_{\text{llr}}$ et $\text{C}_{\text{llr}}^{\text{min}}$ pour les conditions "Specific" et "Random" ($\text{C}_{\text{llr}}$=0.126 et $\text{C}_{\text{llr}}^{\text{min}}$=0.117 pour la référence).

| | $\text{C}_{llr}$ | | $\text{C}_{llr}^{min}$ | | Duration (s) | |
|---|---|---|---|---|---|---|
| **Category** | **Withdrawn** | | **Withdrawn** | | **Mean** | **SD** |
| | **Specific** | **Random** | **Specific** | **Random** | | |
| **NV** | 0.14689 | 0.12941 | 0.13498 | 0.11975 | 3.14 | 1.56 |
| **NC** | 0.13713 | 0.12815 | 0.12728 | 0.11897 | 2.05 | 1.03 |
| **OV** | 0.15396 | 0.14689 | 0.14601 | 0.12819 | 13.00 | 5.50 |
| **L** | 0.12966 | 0.13032 | 0.12173 | 0.12029 | 4.03 | 1.96 |
| **P** | 0.13278 | 0.13431 | 0.12244 | 0.12228 | 7.72 | 3.40 |
| **F** | 0.12703 | 0.13238 | 0.12007 | 0.12135 | 5.84 | 2.68 |

Nous avons ensuite abordé notre analyse par classe phonétique avec le protocole de "knock-out" décrit précédemment. La table 1 montre l'impact des 6 catégories phonémiques sur $\text{C}_{llr}$ pour les conditions "Specific" et "Random"Elle fournit également la quantité de trames de parole pour chaque classe de phonèmes (moyenne et écart par rapport aux extraits de paroles). Une grande variation est observée entre les classes phonémiques : le retrait des voyelles nasales, des consonnes nasales ou des voyelles orales conduit à une perte d'information par rapport à la condition de contrôle ("**Random**") tandis que l'absence de plosives, liquides et fricatives n'a pas d'impact significatif sur la précision du système. Ce résultat est en concordance avec notre revue de la littérature.

Cependant, les fricatives présentent un faible pouvoir discriminatoire, un résultat en contradiction avec la littérature dont, notamment, (Gallardo *et al.*, 2014). La table 2 reprend la même expérience mais en bande large et en se focalisant sur les fricatives et les voyelles orales. Cette fois-ci, les fricatives montrent un pouvoir discriminant supérieur à la moyenne alors que les voyelles orales perdent cette caractéristique. Ce résultat montre clairement -et ce n'est pas réellement une surprise- l'importance de la bande passante. Au vu du contexte "forensique" de ce travail, seuls des résultats obtenus en bande téléphonique sont présentés dans la suite de ce document.

TABLE 2 – Valeurs de $C_{llr}$ et $C_{llr}^{min}$ pour les conditions “Specific” et “Random” pour les fricatives et les voyelles orales en bande large

| Category | $C_{llr}$ | | $C_{llr}^{min}$ | | Duration (s) | |
|---|---|---|---|---|---|---|
| | Withdrawn | | Withdrawn | | Mean | SD |
| | Specific | Random | Specific | Random | | |
| **F** | 0.11738 | 0.11334 | 0.10713 | 0.10394 | 5.84 | 2.68 |
| **OV** | 0.11845 | 0.12216 | 0.11127 | 0.10824 | 13.00 | 5.50 |

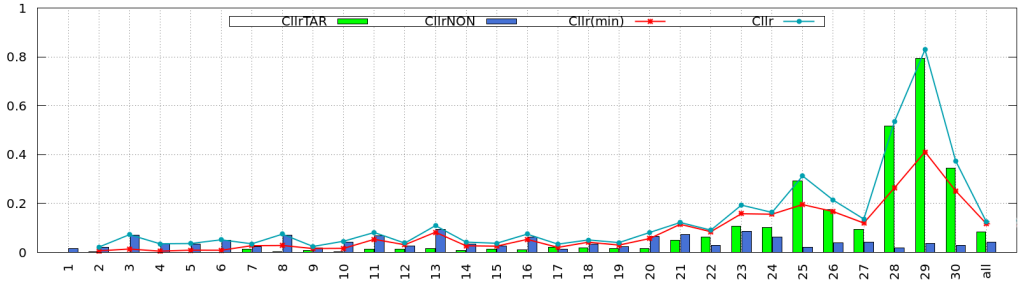



FIGURE 1 – $C_{llr}$, $C_{llr}^{min}$, $C_{llr}^{TAR}$, $C_{llr}^{NON}$ *par locuteur et pour “all” (les données de tous les locuteurs sont utilisées).*

Nous poursuivons notre analyse en réalisant un focus par locuteur et en fonction de la catégorie des comparaisons, cible ou imposteur. La figure 1 résume les résultats.Elle montre que la perte d’information liée aux comparaisons non-cibles (mesurée par $C_{llr}^{NON}$) présente une variation assez faible en fonction du locuteur alors que la variation est importante pour les comparaisons cibles (mesurée par $C_{llr}^{TAR}$). Notons également que la perte d’information provenant des essais cibles (calculée par $C_{llr}^{TAR}$) est principalement responsable des coûts élevés attachés à certains locuteurs.

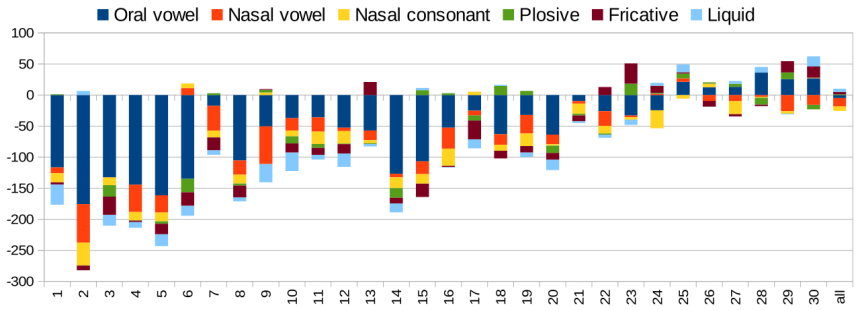



FIGURE 2 – $C_{llr}^{R}$ *par locuteur et “all”.*

La figure 2 présente la contribution de chaque classe phonétique au $C_{llr}^{R}$ en fonction du locuteur. La tendance générale rejoint les résultats présentés dans la table 1 mais une grande variabilité des résultats par classes phonétiques en fonction du locuteur considéré est à noter. Par exemple, le locuteur 2 montre une perte relative de 175% lorsque les voyelles orales sont supprimées quand le 28 accepte un gain de 40% dans la même situation.

La figure 3 montre l’impact de chaque classe phonétique en termes de $C_{llr}^{NON}$, le $C_{llr}^{R}$ relatif calculé sur $C_{llr}^{NON}$ en utilisant les seules paires non-target. Cet indice est supposé être lié principalement au pouvoir de discrimination entre les locuteurs. Les 6 classes phonétiques apparaissent porteuses de pouvoir de discrimination. Les voyelles orales arrivent en premier en termes de pouvoir de discrimination, avec une large avance sur les classes suivantes. Les nasales, voyelles en tête suivies par les consonnes, prennent les places suivantes. Les classes restantes ont un comportement similaire bien que porteuses de moins d’information discriminante du locuteur. Les résultats obtenus sont plutôt consistent pour les 30 locuteurs étudiés.

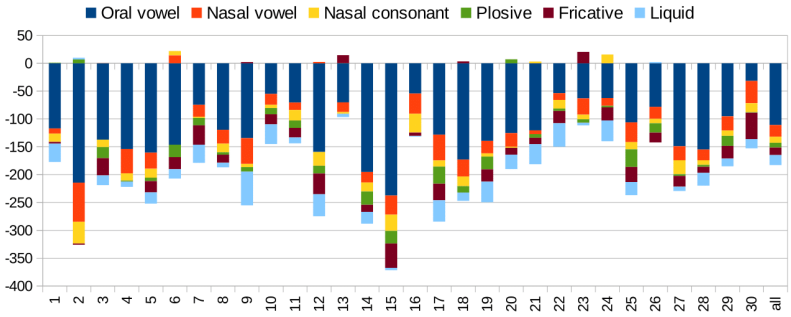



FIGURE 3 – $C_{\text{llr}}^{\text{R}}$ *calculé sur* $C_{\text{llr}}^{\text{NON}}$ *par locuteur et "all".*

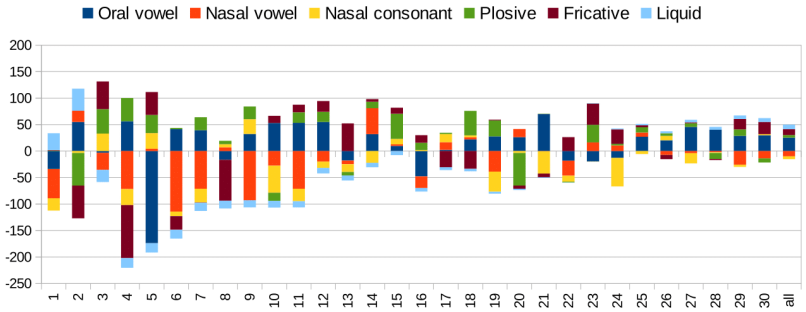



FIGURE 4 – $C_{\text{llr}}^{\text{R}}$ calculé sur $C_{\text{llr}}^{\text{TAR}}$ par locuteur et "all".

La figure 4 utilise un principe similaire à la figure 3. Elle présente l'impact des classes phonémiques par locuteur, en termes de $C_{\text{llr}}^{\text{TAR}}$. $C_{\text{llr}}^{\text{TAR}}$ est calculé uniquement en utilisant des comparaisons cibles et, grâce à notre protocole offrant un grand nombre par locuteur de ces tests, traduit directement les effets de la variabilité intra-locuteur. En opposition avec les résultats précédents, pour ce $C_{\text{llr}}^{\text{TAR}}$, retirer les voyelles orales des enregistrements conduit à une amélioration de $C_{\text{llr}}$ pour environ $70\%$ des locuteurs. Ces sons semblent donc, ici, apporter peu en termes de discrimination du locuteur ou même perturber la discrimination. Les classes fricatives, liquides et plosives ont le même comportement que les voyelles orales. Au contraire, les nasales (et en particulier les voyelles nasales) jouent un rôle positif : retirer ces phonèmes augmente le $C_{\text{llr}}$.

Le rôle positif des nasales pour la comparaison des locuteurs pourrait s'expliquer par la contribution importante des cavités nasales et para-nasales. Cet aspect morphologique propre à ces phonèmes constitue un élément que les locuteurs peuvent difficilement contrôler (volontairement ou involontairement). Cela induit une faible variabilité intra-locuteur, pour une variabilité inter-locuteur significative. (Stevens, 1999; Schindler & Draxler, 2013).

Les voyelles orales apportent la plus grande partie en termes de pouvoir de discrimination des locuteurs mais présente en même temps une grande variabilité intra-locuteur, liée à une part significative des pertes de performance. La variabilité intra-locuteur apporte en général environ deux tiers des pertes de $C_{llr}$ (0,66 contre 0,33 pour les pertes portées par la variabilité inter-locuteur). Cette proportion est significativement plus élevée (jusqu'à 0,94 vs 0,06) pour les locuteurs qui présentent la plus grande contribution à la perte de $C_{llr}$. Il est intéressant de lier cette constatation à deux faits : (1) presque toutes les classes de phonèmes étudiées aident à la discrimination des locuteurs pour tous les locuteurs ; (2) certaines classes de phonèmes dégradent la partie cible de $C_{llr}$ lorsque d'autres classes offrent un comportement positif. Il est intéressant de remarquer que la même classe de phonèmes peut avoir un comportement très différent selon le locuteur, ce qui renforce la nécessité d'une prise en compte fine de l'effet locuteur.

# 5 Conclusion

Cet article est consacré à l'étude de impact du contenu phonémique sur le processus de comparaison vocale. Pour cela, il utilise un système automatique de reconnaissance du locuteur comme instrument de mesure et un protocole de "knock-out" et une bande téléphonique appropriée au contexte criminalistique de cette étude.

Nous avons étudié l'impact de chaque classe phonémique sur la performance de la comparaison vocale mesurée avec le critère $C_{llr}$. Les résultats ont montré que toutes les classes phonémiques jouent un rôle en termes de pouvoir de discrimination du locuteur. Les voyelles orales, les voyelles nasales et les consonnes nasales, dans cet ordre, sont meilleures que le contenu phonémique moyen en termes de performance de comparaison vocale, rejoignant des constats précédents. Les fricatives, par contre, n'apportent pas plus qu'un contenu moyen. Ce résultat surprenant par rapport à la littérature a été expliqué par le choix d'une bande passante étroite : en bande large, cette catégorie retrouve sa pertinence connue en termes de discrimination des locuteurs.

Lorsque nous nous sommes concentrés sur la variabilité intra-locuteur, les voyelles orales sont apparues liées à un niveau élevé de $C_{llr}$ intra-locuteur. Nous avons vu précédemment que cette classe phonémique apportait une grande partie du pouvoir de discrimination du locuteur mais elle apparaît également très sensible à la variabilité intra-locuteur. En revanche, les nasales ont montré une bonne capacité de discrimination et, en même temps, apparaissent robustes en ce qui concerne la variabilité intra-locuteur.

Dans cet article, nous avons souligné à plusieurs reprises l'importance du facteur locuteur. Nous avons observé de grandes variations de $C_{llr}$ et de $C_{llr}^{TAR}$ entre nos 30 locuteurs. Nous avons également observé en fonction du locuteur des comportements très différents du système en termes de $C_{llr}^{TAR}$ suivant les classes phonémiques sélectionnées.

La principale conclusion du travail présenté est une remise en cause nécessaire des protocoles d'évaluation habituels en reconnaissance automatique du locuteur (ASpR). Ces protocoles se basent en effet principalement sur la discrimination des locuteurs ($C_{llr}^{NON}$) en négligeant largement la variabilité intra-locuteur ($C_{llr}^{TAR}$) Il nous apparaît obligatoire de prendre en considération de manière approfondie la variabilité intra-locuteur ainsi que le facteur locuteur lui-même. Cela est particulièrement important lorsqu'il s'agit d'attester de la fiabilité d'une solution de comparaison de voix dans le domaine criminalistique.

Les résultats présentés dans cet article restent préliminaires (100 extraits de parole par locuteur, peu de variabilité contextuelle et seulement 30 locuteurs masculins, français natifs). Dans nos futurs travaux, nous souhaitons effectuer une analyse similaire sur une base de données supérieure d'un ordre de magnitude (~1000 enregistrements par locuteurs et plusieurs centaines de locuteurs). Une telle base permettrait également d'affiner notre étude en nous intéressant aux phonèmes individuels au lieu de classes phonémiques. Enfin, ce travail nous permet de promouvoir un système automatique de comparaison de voix **explicite** et **transparent** capable d'analyser en profondeur le contenu phonétique des extraits de discours et de détailler ses sorties dans un langage compréhensible par un expert. Loin d'être un simple rêve, de telles caractéristiques sont certainement un élément indispensable pour une utilisation responsable de systèmes de comparaison de voix en milieu criminalistique.

# 6 Remerciements

La recherche rapportée ici a été soutenue par le projet ANR-12-BS03-0011 FABIOLE.

# Références


AITKEN C. G. & TARONI F. (2004). *Statistics and the evaluation of evidence for forensic scientists*, volume 10. Wiley Online Library.

AJILI M., BONASTRE J., KAHN J., ROSSATO S. & BERNARD G. (2016a). Fabiole, a speech database for forensic speaker comparison. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*

AJILI M., BONASTRE J.-F., BEN KHEDER W., ROSSATO S. & KAHN J. (2016b). Phonetic content impact on forensic voice comparison. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*, p. 210–217 : IEEE.

AJILI M., F. BONASTRE J., ROSSATTO S. & KAHN J. (2016c). Inter-speaker variability in forensic voice comparison : A preliminary evaluation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 2114–2118.

AMINO K., OSANAI T., KAMADA T., MAKINAE H. & ARAI T. (2012). Effects of the phonological contents and transmission channels on forensic speaker recognition. In *Forensic Speaker Recognition*, p. 275–308. Springer.

AMINO K., SUGAWARA T. & ARAI T. (2006). Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties. *Acoustical science and technology*, **27**(4), 233–235.

ANTAL M. & TODEREAN G. (2006). Speaker recognition and broad phonetic groups. In *SPPRA*, p. 155–159.

BESACIER L., BONASTRE J.-F. & FREDOUILLE C. (2000). Localization and selection of speaker-specific information with statistical modeling. *Speech Communication*, **31**(2), 89–106.

BONASTRE J.-F., SCHEFFER N., MATROUF D., FREDOUILLE C., LARCHER A., PRETI A., POUCHOULIN G., EVANS N. W., FAUVE B. G. & MASON J. S. (2008). Alize/spkdet : a state-of-the-art open source software for speaker recognition. In *Odyssey*, p.20.

BONASTRE J.-F., WILS F. & MEIGNIER S. (2005). Alize, a free toolkit for speaker recognition. In *ICASSP (1)*, p. 737–740.

BRÜMMER N., BURGET L., ČERNOCKỲ J. H., GLEMBEK O., GREZL F., KARAFIAT M., VAN LEEUWEN D. A., MATĚ P., SCHWARZ P. & STRASHEIM A. (2007). Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006. *Audio, Speech, and Language Processing, IEEE Transactions on*, **15**(7), 2072–2084.

BRÜMMER N. & DU PREEZ J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, **20**(2), 230–275.

CASTRO D. R. (2007). *Forensic evaluation of the evidence using automatic speaker recognition systems*. PhD thesis, Universidad autónoma de Madrid.

CHAMPOD C. & MEUWLY D. (2000). The inference of identity in forensic speaker recognition. *Speech Communication*, **31**(2), 193–203.

DEHAK N., KENNY P., DEHAK R., DUMOUCHEL P. & OUELLET P. (2011). Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, **19**(4), 788–798.

EATOCK J. P. & MASON J. S. (1994). A quantitative assessment of the relative speaker discriminating properties of phonemes. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, volume 1, p. I–133 : IEEE.

GALLARDO L. F., WAGNER M. & MÖLLER S. (2014). I-vector speaker verification based on phonetic information under transmission channel effects. In *INTERSPEECH*, p. 696–700.

GALLIANO S., GEOFFROIS E., MOSTEFA D., CHOUKRI K., BONASTRE J.-F. & GRAVIER G. (2005). The ester phase ii evaluation campaign for the rich transcription of french broadcast news. In *European Conference on Speech Communication and Technology*, p. 1149–1152.

GIRAUDEL A., CARRÉ M., MAPELLI V., KAHN J., GALIBERT O. & QUINTARD L. (2012). The repere corpus : a multimodal corpus for person recognition. In *LREC*, p. 1102–1107.

GONZALEZ-RODRIGUEZ J. & RAMOS D. (2007). Forensic automatic speaker classification in the "coming paradigm shift". In *Speaker Classification I*, p. 205–217. Springer.

GRAVIER G., ADDA G., PAULSON N., CARRÉ M., GIRAUDEL A., GALIBERT O. *et al.* (2012). The etape corpus for the evaluation of speech-based tv content processing in the french language. *International Conference on Language Resources, Evaluation and Corpora*.

HOFKER U. (1977). Auros-automatic recognition of speakers by computers : phoneme ordering for speaker recognition. In *Proc. 9th International Congress on'Acoustics, Madrid*, p. 506–507.

KASHYAP R. (1976). Speaker recognition from an unknown utterance and speaker-speech interaction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **24**(6), 481–488.

LARCHER A., BONASTRE J.-F., FAUVE B. G., LEE K.-A., LÉVY C., LI H., MASON J. S. & PARFAIT J.-Y. (2013). Alize 3.0-open source toolkit for state-of-the-art speaker recognition. In *INTERSPEECH*, p. 2768–2772.

LINARES G., NOCÉRA P., MASSONIE D. & MATROUF D. (2007). The lia speech recognition system : from 10xrt to 1xrt. In *International Conference on Text, Speech and Dialogue*, p. 302–308 : Springer.

MAGRIN-CHAGNOLLEAU I., BONASTRE J.-F. & BIMBOT F. (1995). Effect of utterance duration and phonetic content on speaker identification usind second order statistical methods. In *Proceedings of EUROSPEECH*.

MATROUF D., SCHEFFER N., FAUVE B. G. & BONASTRE J.-F. (2007). A straightforward and efficient implementation of the factor analysis model for speaker verification. In *INTERSPEECH*, p. 1242–1245.

MORRISON G. S. (2009). Forensic voice comparison and the paradigm shift. *Science & Justice*, **49**(4), 298–308.

PRINCE S. J. & ELDER J. H. (2007). Probabilistic linear discriminant analysis for inferences about identity. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, p. 1–8 : IEEE.

PROVIDERS A. (2009). Standards for the formulation of evaluative forensic science expert opinion. *Sci. Justice*, **49**, 161–164.

SAMBUR M. (1975). Selection of acoustic features for speaker identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **23**(2), 176–182.

SCHINDLER C. & DRAXLER C. (2013). The influence of bandwidth limitation on the speaker discriminating potential of nasals and fricatives. *International Association for Forensic Phonetics and Acoustics (IAFPA)*.

STEVENS K. (1999). Acoustic phonetics. 1998.

WOLF J. J. (1972). Efficient acoustic parameters for speaker recognition. *The Journal of the Acoustical Society of America*, **51**(6B), 2044–2056.

# Suivre le rythme de tes paroles

Solange Rossato[1], Dan Zhang[1], Moez Ajili[2], Jean-François Bonastre[2]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France
(2) Univ. Avignon et Pays de Vaucluse, LIA, F-84000 Avignon France
solange.rossato@univ-grenoble-alpes.fr

## RÉSUMÉ

Différentes mesures temporelles, telles que la durée des voyelles et des consonnes, ont été proposées pour tenter de caractériser le rythme de la parole et classer ainsi les langues, les dialectes ou les idiolectes. C'est sur ce dernier rôle des paramètres temporels de la parole que cette étude se focalise en s'appuyant sur la base de données FABIOLE. Utilisée pour la comparaison de voix, elle est construite à partir d'émissions médiatiques (TV et radio). Elle nous permet ainsi d'étudier la variabilité de certains paramètres temporels, variabilité intra et inter locuteurs, à la recherche d'un idiolecte. Les résultats montrent que la part de variabilité que l'on peut attribuer au locuteur atteint 45% pour la variance de la durée des segments non voisés, 42% pour le pourcentage total de segments voisés. Ainsi, ces mesures temporelles dépendent du locuteur, de façon bien plus marquée que ne le sont les paramètres formantiques.

## ABSTRACT

**Following the rhythm of your speech**
Various temporal measures, such as the duration of vowels and consonants, have been proposed to characterize the rhythm of speech and thus classify languages, dialects or idiotic expressions. It is on this last role of the temporal parameters of speech that this study focuses on, using the FABIOLE database. Used for voice comparison, it is constructed from media broadcasts (TV and radio). It allows us to study the variability of certain temporal parameters, within and between speakers, in search of idiosyncrasy. The results show that the percentage of variability that can be attributed to the speaker is 45% for the variance of the duration of un-voiced segments, 42% for the total percentage of voiced segments. Thus, these temporal measurements depend on the speaker, much more strongly than the formantic parameters.

MOTS-CLÉS : Mesures temporelles, rythme, voisement, locuteur, idiolect, idiosyncrasie
KEYWORDS: Timing, rythm, voicing, speaker, idiolect, idiosyncrasy

# 1 Parole et Identité : le choix de mesures rythmiques

Le rythme est une notion de structuration temporelle du flux de parole. Ainsi, une typologie ryhtmique classe les langues comme étant plutôt syllabiques, accentuelles ou moraiques. Les mesures ryhtmiques, quelles soient basées sur l'alternance Consonne Voyelle (Ramus, 1999) ou sur des alternances d'intervalles voisés et non voisées ou des intervalles entre syllabes accentuées (Dellwo & Fourcin, 2013a; Dellwo, Fourcin, & Abberton, 2007) montrent en effet un impact de la langue parlée important sans pour autant permettre une réelle classification typologique. En effet, d'autres facteurs entrent en jeu, tel que le style de parole ou phonostyle (de Mareüil, 2014; Fónagy,


10.21437/JEP.2018-5

1983; Simon, Auchlin, Avanzi, & Goldman, 2010) qui modifient l'organisation temporelle de la parole. Il a été ainsi montré que les pauses jouent un rôle spécifique dans la parole politique (Duez, 1999). (Eskénazi, 1993) écrit que le style de parole reflète une interaction entre un locuteur et son environnement : « *It is the perception of the various status levels of his listener and of the type of situation in which he finds himself.* » (p. 502). La situation de communication est donc un facteur important mais le rôle central du locuteur, ses projections et son histoire est également souligné par l'auteur et rejoint la définition de (Labov, 1972) pour qui le style de parole est un fait qui relève de l'individualité du locuteur. Le caractère idiosyncrasique des paramètres rythmiques a ainsi été mis en évidence par plusieurs études (Dellwo, Leemann, & Kolly, 2012, 2015; Leemann, Kolly, & Dellwo, 2014). Le rythme de la parole apparaît comme un élément dépendant du locuteur. Nous allons alors suivre le rythme des paroles de locuteurs pour tenter de quantifier la relation entre ryhtme et l'identité du locuteur. Pour cela, la base de données FABIOLE, contenant plus de 30h de parole médiatique, pour 30 locuteurs différents, soit 100 extraits de parole par locuteur, est un corpus contenant une grande variabilité intra- et inter- locuteurs. L'objectif de cette étude est de commencer par des paramètres rythmiques concernant le voisement afin de déterminer dans quelle mesure ces paramètres temporels nous renseignent sur la provenance de la voix. Après avoir détaillé la méthodologie mise en œuvre, les résultats seront présentés avant la partie discussion.

# 2 Méthodologie

Nous avons utilisé une approche basée sur un grand corpus de parole contenant environ 30h de parole, le corpus FABIOLE. Cette approche nous a conduit à envisager des mesures automatiques, ne nécéssitant pas ou un minimum d'interventions manuelles. Les mesures rythmiques qui en découlent sont donc d'un très bas niveau, et se limitent ici à une analyse grossière des alternances de parties voisées et non voisées dans les extraits de parole. Il ne s'agit en aucun cas d'une analyse fine, et nous sommes bien loin des mesures rythmiques s'appuyant sur les syllabes accentués. L'avantage de ce type d'analyse bas niveau est qu'elle permet une extraction de paramètres rythmiques qui peut se faire de façon automatique mais sur un très grand nombre d'extraits de parole.

Par ailleurs, nous n'avons pas de transcription orthographique pour l'intégralité de la base de données. Plutôt que de se baser sur une transcription obtenue par un système de transcription automatique (Ajili, 2017), nous avons pris le parti de nous appuyer sur des mesures ne nécessitant aucune transcription préalable, effectuées directement sur le signal acoustique sans faire appel à d'autres ressources. Les indices de voisement se prètent parfaitement à cela. Bien évidemment, ces indices dépendent de la phonotactique de la langue, mais ce facteur est neutralisé dans notre base de données contenant exclusivement des extraits de locuteurs francophones.

## 2.1 La base de données FABIOLE

La base de données utilisée est celle développée dans le cadre du projet ANR Fiabilité en biométrique vocale. Ce corpus FABIOLE (Ajili, Bonastre, Kahn, Rossato, & Bernard, 2016) a été extrait de programmes télévisuels ou radiophoniques francophones entre 2013 et 2014 avec l'objectif de capturer les variabilités de la parole intra- et inter- locuteurs dans un contexte de comparaison de voix. Le corpus s'est focalisé sur les voix d'hommes en raison de leur plus grande disponibilité dans les médias. Ainsi, un minimum de 100 extraits de minimum 30 sec chacun est obtenu pour 30 locuteurs différents. La base de données FABIOLE contient également 100 extraits de 30 sec produits par 100 locuteurs différents mais cette partie du corpus n'est pas pris en compte ici. Les variabilités dues canal de communication sont réduites. En effet, les extraits de parole sont généralement de très bonne qualité, étant enregistré avec du matériel audio professionnel. Ainsi, les

analyses ont pu être menées sur l'ensemble du corpus excepté pour un locuteur, chroniqueur, dont les enregistrements n'ont pas lieu en studio et qui sera par la suite enlevé de l'analyse. Certains signaux contiennent de la musique de fond.

Les locuteurs ont des métiers différents : journalistes, hommes politiques, chroniqueurs mais étant donné leur grande présence dans les médias ont tous une très grande pratique de la parole médiatique. Le diagramme de la Figure 1 présente cette répartition et l'on note clairement la grande proportion de journalistes, chroniqueurs et débatteurs. Cependant, plutôt que de considérer le métier des locuteurs, nous nous sommes intéressés à la situation de la communication, en prenant en compte les différentes émissions dont sont extraits les signaux de parole de la base FABIOLE comme autant de situations différentes.

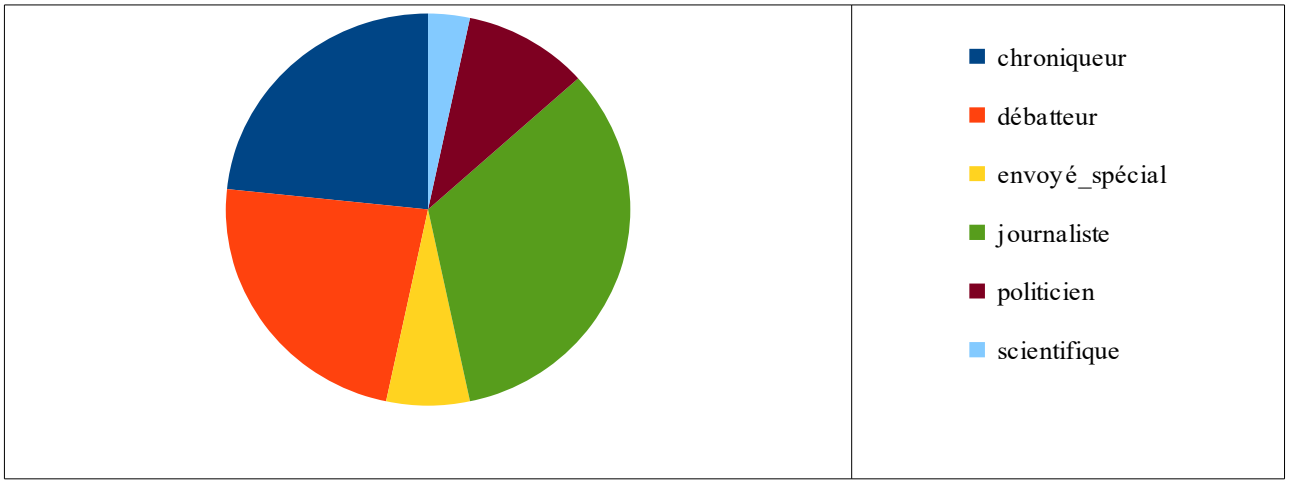



FIGURE 1: Diagramme de répartition des métiers des locuteurs

Ces extraits de parole proviennent de 9 émissions :

- des débats : "Ça vous regarde - Le débat" (cvgdde-bat), "Entre les Lignes" (entreligne), "Le masque et la plume" (MsqPlum),
- des chroniques : "Service public" (Spublic), "Comme on nous parle" (ComParle),
- des séances parlementaires : "Top Questions" (topquestions),
- des journeaux télévisés "BFM Story" (bfmstory), "LCP Info"(parlinfo), "Ca vous regarde-l'Info (cvgdinfo)".

La répartition entre les différentes émissions n'est pas uniforme, ainsi que le montre la Figure 2. Par rapport à (Ajili et al., 2016), il manque une émission qui correspond au locuteur que nous avons écarté pour des raisons de qualité acoustique des signaux. Il y a ainsi un fort lien entre locuteurs et émissions, certains locuteurs étant les animateurs des émissions. Ainsi, 19 locuteurs sur les 29 n'interviennent très majoritairement (plus de 90 % des extraits) que dans une seule émission. Les locuteurs interviennent dans plusieurs émissions mais avec des temps de parole très disparates. Il n'est pas possible, avec ces données, d'étudier conjointement l'influence du locuteur et de l'émission. Or ces deux facteurs influencent le ryhtme de la parole. Nous allons donc les étudier de façon indépendante tout en étant conscients que c'est une limite forte que de ne pas pouvoir étudier l'interaction entre ces deux facteurs.

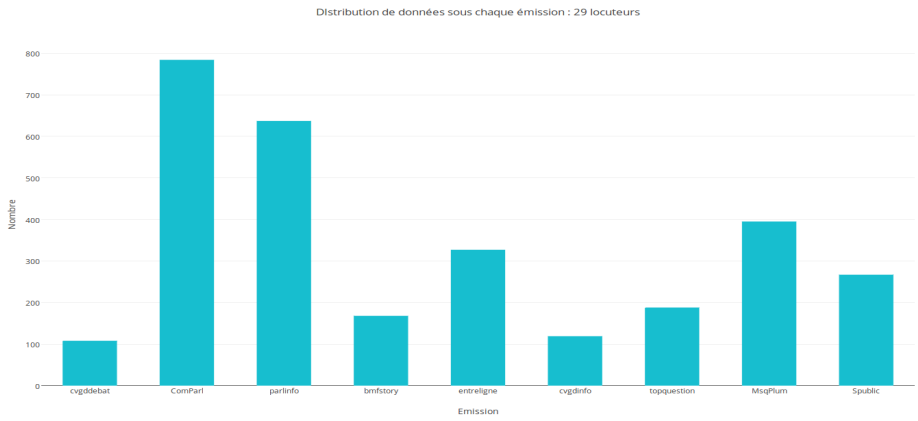



FIGURE 2: Répartition des extraits de parole en fonction des émissions

## 2.2 Les mesures temporelles du voisement

Ce premier travail s'est focalisé sur un premier aspect rythmique de la parole, à savoir l'alternance des intervalles voisés et de intervalles non voisés (incluant les pauses). Ce travail s'inspire largement des mesures effectuées des études de (Dellwo & Fourcin, 2013b; Dellwo et al., 2012, 2015; Leemann et al., 2014) Nous avons utilisé le script de ProsodyPro[1] développé par (Xu, 2013) pour automatiser la mesure du voisement. A partir des mesures de f0, nous avons obtenu les durées des intervalles voisés *dVO* et non voisés *dUV* pour chacun extrait. Pour chaque fichier, nous avons calculé les mesures suivantes :

- le pourcentage global de la durée cumulée des intervalles voisés dans l'extrait sonore, *%VO ;*
- la durée moyenne des intervalles voisés, *Moyenne(DuréeVO)*;
- le coefficient de variation de la durée des intervalles non voisés, *VarcoUV*, calculé de la façon suivante : 100*Var(*dUV*)/Moy(*dUV*)
- le coefficient de variation de la durée des intervalles voisés, *VarcoVO*.

Ainsi, un extrait de parole qui contient des pauses importantes aurait un *%VO* légèrement plus faible et un coefficient de variance *VarcoUV* plus grand que celui qui contient moins de longues pauses. A ces mesures classiques, nous avons également ajouté les moyenne et coefficient de variation de l'intervalle de temps entre le début de deux intervalles voisés successifs ou *paire*, le dernier intervalle voisé n'étant pas pris en compte. La *paire* est ainsi l'intervalle de durée $dVO_i+dUV_{i+1}$. Parmi ces *paires*, nous avons cherché la proportion de *paires* pour lesquelles l'intervalle voisé $dVO_i$ est plus court que l'intervalle non voisé $dUV_{i+1}$. A la liste précédente, viennent donc s'ajouter :

- le proportion de *paires* pour lesquelles la durée d'un intervalle non voisé est supérieure à celle de l'intervalle voisé qui le précède, $\%(UV_{i+1} > VO_i)$ ;
- la durée moyenne de chaque *paire* d'intervalles voisés et non voisés, *Moyenne(paire)*;
- le coefficient de variation de la durée des *paires*, *VarcoPaire*.

La Figure 3 permet de visualiser les durées extraites du signal et permettant de calculer ces 7 paramètres temporels de voisement.

[1] http://www.homepages.ucl.ac.uk/~uclyyix/ProsodyPro/

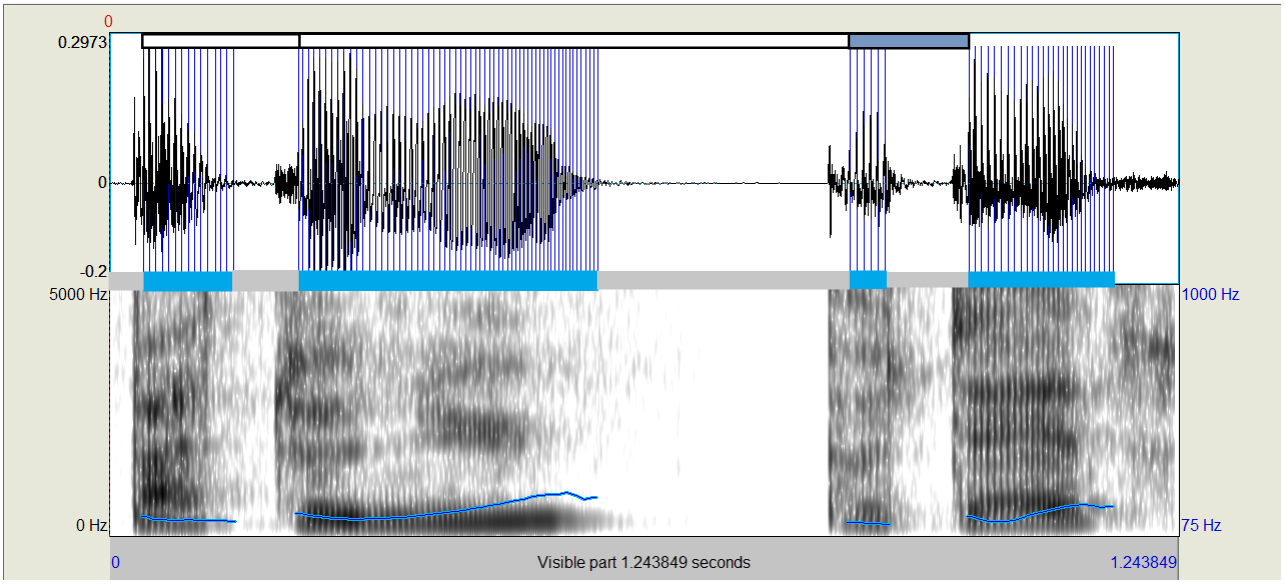



FIGURE 3: Signal de parole et sonagramme avec, au dessus du sonagramme, la visualisation des intervalles voisés (en bleu) et non voisés (en gris) et, au dessus du signal de parole, la visualisation des paires (encadrés noirs) avec en bleu, celle pour laquelle la portion voisée est plus courte que la portion non voisée qui la suit.

### 2.3 Analyses statistiques

Des ANOVA à un facteur, Locuteur ou Émission, sont appliquées sur chaque variable temporelle de façon indépendante. Les résultats sont significatifs étant donné le grand nombre de valeurs (avec 100 valeurs par locuteur). Pour quantifier l'influence d'un facteur sur chaque variable, nous avons calculé la taille de l'effet en utilisant l'êta-carré $\eta^2$. L'êta-carré décrit la force de la relation, et correspond au pourcentage de la variance totale expliquée par le facteur en question. Plus cette valeur est importante, plus le facteur en question est explicatif de la variance de la variable étudiée. Une interprétation fréquente (Cohen, 1988) indique une taille de l'effet faible lorsque la valeur est inférieure à 1 %, moyenne jusqu'à 6 %, et importante lorsque les valeurs sont supérieures à 14 %. La formule de calcul de l'éta-carré est la suivante, avec *SStotal* étant la variance totale et *SSbetween*, la variance des moyennes par locuteur :

$$\eta^2 = \frac{SSbetween}{SStotal} * 100$$

Après avoir présenté les résultats sous forme de graphiques permettant de visualiser la variation intra- et inter- locuteurs, les résultats de l'analyse statistique sont présentés.

## 3 Résultats

### 3.1 Un rythme spécifique au locuteur ?

L'intégralité des mesures obtenues sont disponible dans le fichier .odt téléchargeable avec l'article. Les graphiques et analyses statistiques ont été réalisés avec Matlab®. La Figure 4 montre la

répartition du pourcentage de voisement *%VO* en fonction des locuteurs. Globalement, sur l'ensemble des extraits, 60,2 % du signal est voisé, cette proportion moyenne variant de 53,7 % (loc. 6) à 65,5 % (loc. 19) en fonction des locuteurs.

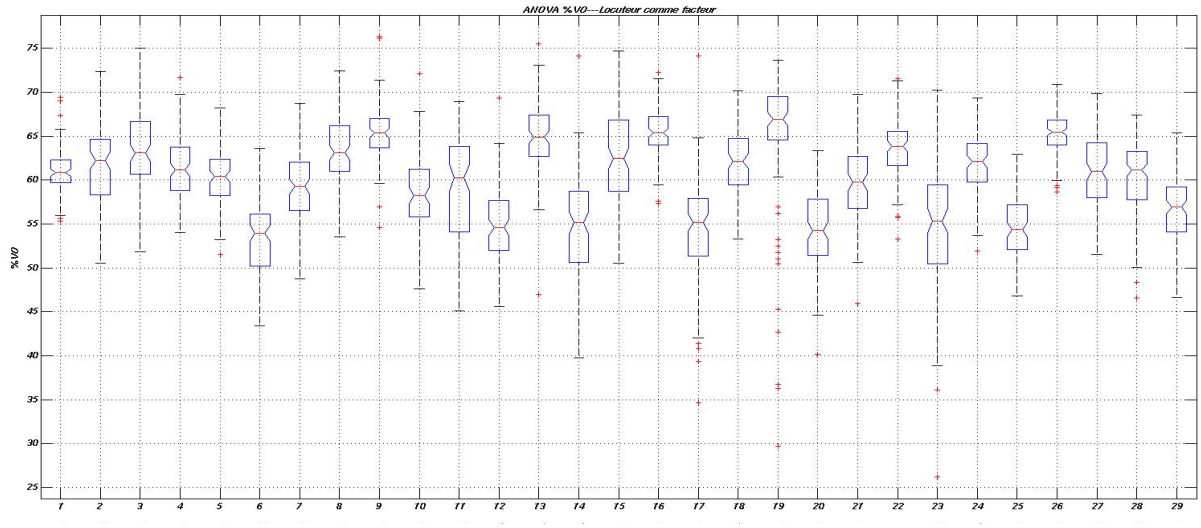



FIGURE 4: Répartition (médiane et quartile) du pourcentage de voisement %VO en fonction des locuteurs.

Une ANOVA à un facteur est effectuée pour chacune des 7 variables. Les résultats des ANOVA étudiant le facteur locuteur, tous significatifs, ainsi que les valeurs des êta-carré $\eta^2$ sont indiqués dans le tableau 5. Le facteur Locuteur a un effet important pour toutes les variables temporelles étudiées, avec des valeurs largement supérieures au seuil de 14 %. Pour certaines variables, la taille de l'effet est même très importante. Représenter les tailles sous forme d'un radar-chart permet de bien visualiser les variables la force de la relation entre le facteur Locuteur et chacune des variables (voir Figure 6). Il apparaît ainsi clairement que les variables qui montrent le plus de variation propre au locuteur sont la proportion de voisement (*%VO*), le coefficient de variance des intervalles non voisés (*VarcoUV*) ainsi que la durée moyenne des intervalles séparant le début de deux intervalles de voisement consécutifs (*Moyenne(paire)*).

| *Variables* | *Résultats de l'ANOVA* | $\eta^2$ |
|---|---|---|
| %VO | F(28,2964) = 76.83, p <.001 | **42.056** |
| Moyenne(DuréeVO) | F(28,2964) = 64.13, p <.001 | **37.727** |
| VarcoVO | F(28,2964) = 36.22, p <.001 | **25.492** |
| VarcoUV | F(28,2964) = 88.32, p <.001 | **45.485** |
| $\%(UV_{i+1} > VO_i)$ | F(28,2964) = 60.02, p <.001 | **37.686** |
| Moyenne(paire) | F(28,2964) = 73.99, p <.001 | **41.142** |
| VarcoPaire | F(28,2964) = 25.47, p <.001 | **19.391** |

TABLE 5 : Résultats de l'ANOVA étudiant le facteur Locuteur

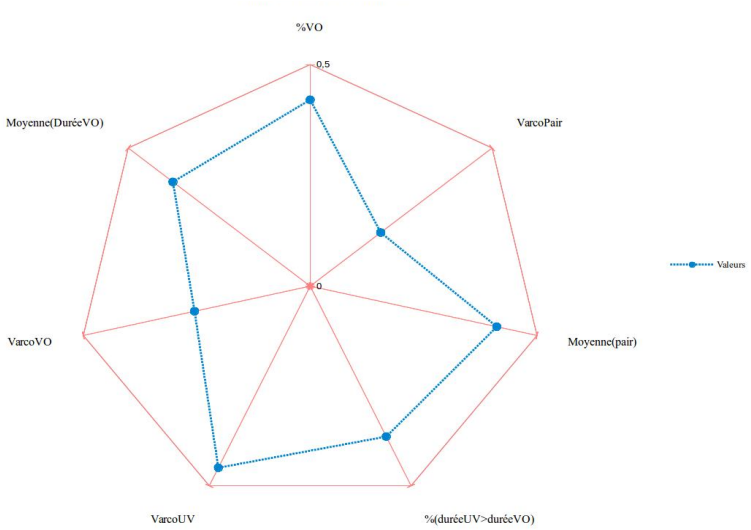



FIGURE 6: Radar-chart des tailles de l'effet Locuteur $\eta^2$ pour chaque variable.

### 3.2 Ou un rythme lié à l'émission médiatique ?

En suivant la même démarche pour les émissions que pour les locuteurs, nous obtenons également une influence significative du facteur Émission sur chaque variable étudiée. Dans l'ensemble, les tailles de l'effet sont plus faibles que celles obtenues pour le facteur Locuteur mais non négligeables ainsi que le montre le tableau 7. On observe ainsi que les variables qui mesurent la proportion de voisement *%VO* ainsi que le coefficient de variance des intervalles non voisés *VarcoUV* sont fortement dépendantes de l'émission.

| *Variables* | *Résultats de l'anova* | $\eta^2$ |
|---|---|---|
| %VO | F(8,891) = 36.56, p<.001 | **24.71** |
| Moyenne(DuréeVO) | F(8,891) = 16.14, p<.001 | *12.66* |
| VarcoVO | F(8,891) = 30.77, p<.001 | **21.56** |
| VarcoUV | F(8,891) = 42.89, p<.001 | **27.81** |
| $\%(UV_{i+1} > VO_i)$ | F(8,891) = 27.99, p<.001 | **20.08** |
| Moyenne(paire) | F(8,891) = 29.36, p<.001 | **20.86** |
| VarcoPaire | F(8,891) = 4.07 , p<.001 | 3.52 |

TABLE 7 : Résultats de l'ANOVA étudiant le facteur Emission

# 4 Discussion et conclusion

L'organisation temporelle du voisement dépend clairement du locuteur, et ce lien de dépendance est très important. En effet, les mesures formantiques des voyelles orales, très largement étudiés comme

apportant des informations idiosyncratiques, ont été étudiées sur ces mêmes données FABIOLE, (Ajili, 2017) et les tailles de l'effet obtenues sont bien plus faibles pour les formants, y compris le F4 que pour les mesures temporelles étudiées ici, pour lesquelles la taille de l'effet varie entre 37,7 % et 45,5 % pour 5 d'entre elles, tandis qu'elle atteint difficilement 20 % pour les valeurs les plus importantes obtenues pour les formants des voyelles orales (voir le tableau 8).

| Vowel | F1 | F2 | F3 | F4 |
|---|---|---|---|---|
| /i/ | 1.65 | *6.16* | *8.09* | **14.08** |
| /y/ | 2.98 | 5.95 | 5.91 | *11.68* |
| /u/ | 2.1 | 2.50 | *6.51* | 3.98 |
| /e/ | 1.83 | **17.46** | *9.72* | **20.56** |
| /ø/ | 5.79 | *7.9* | 4.13 | **14.86** |
| /o/ | *12.2* | *13.22* | *8.1* | 8.10 |
| /ɛ/ | 2.8 | *11.30* | *10.75* | **18.48** |
| /œ/ | *10.88* | *7.60* | *8.48* | **23.04** |
| /ɔ/ | *12.51* | *10.56* | *7.34* | *13.18* |
| /a/ | *12.85* | 4.0 | *13.21* | **19.82** |

TABLE 8 : Tailles de l'effet $\eta^2$ pour les formants des voyelles orales sur la base de données FABIOLE (extrait de (Ajili, 2017) p 154).

Cependant, le rôle de l'émission n'est pas négligeable, loin de là, avec des tailles de l'effet allant jusqu'à 27.8 % pour le coefficient de variance des intervalles non voisés *VarcoUV*. Quelle part, dans l'organisation temporelle globale du voisement, peut-on attribuée au style de parole spécifique du locuteur (son idiolecte) et quelle part peut-on attribuer au style de parole correspondant à la situation de communication ? Ici, les situations de communication ont toutes en commun d'être d'une parole publique et médiatique et tous nos locuteurs sont des professionnels de la parole, mais cette parole médiatique n'a pas les mêmes objectifs lorsqu'il s'agit de journaux télévisés ou de sessions parlementaires, avec un aspect plus ou moins formel. La question de l'interaction entre idiolecte et situation de communication due à l'émission reste ainsi posée. Il faudrait, à l'instar des travaux de Dellwo (Dellwoa & Schmida, 2016) qui étudient l'interaction entre la langue et l'idiolecte en enregistrant des locuteurs bilingues, étudier cette problématique dans son ensemble en croisant locuteurs et situations de communication pour étudier leur influence respective et leur interaction sur les paramètres rythmiques. Peut-on, parmi les multiples mesures rythmiques possibles trouver celles qui relèvent plutôt de tel ou tel facteur ?

Une importante limitation à cette étude est sa focalisation sur des mesures d'organisation temporelle du voisement. Or le rythme ne peut se résumer à cette mesure bas niveau et nous devons compléter cette étude, sur cette même base de données avec une estimation des centres de pseudo-syllabes, une détection des pseudo-syllabes accentuées pour permettre de mesurer des intervalles entre syllabes accentuées.

# Remerciements

Ce travail a pu se faire grâce au financement du projet ANR-12-BS03-0011 FABIOLE.

# Références


Ajili, M. (2017, novembre 28). *Fiabilité de la comparaison des voix dans le cadre judiciaire*. Université d'Avignon et des Pays du Vaucluse, Avignon, France.

Ajili, M., Bonastre, J.-F., Kahn, J., Rossato, S., & Bernard, G. (2016). Fabiole, a speech database for forensic speaker comparison. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC* (p. 23–28).

Cohen, J. (1988). Statistical power analysis for the behavioral sciences second edition. *Lawrence Erlbaum Associates, Publishers*.

de Mareüil, P. B. (2014). Qu'est-ce qu'un (phono) style? *Cahiers de linguistique française*, (31), 9–19.

Dellwo, V., & Fourcin, A. (2013). Rhythmic characteristics of voice between and within languages. *Revue Tranel (Travaux neuchâtelois de linguistique), 59*, 87–107.

Dellwo, V., Fourcin, A., & Abberton, E. (2007). Rhythmical classification of languages based on voice parameters. *Proceedings of ICPhS XVI*, 1129–1132.

Dellwo, V., Leemann, A., & Kolly, M.-J. (2012). Speaker idiosyncratic rhythmic features in the speech signal. In *INTERSPEECH* (p. 1584–1587).

Dellwo, V., Leemann, A., & Kolly, M.-J. (2015). Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *The Journal of the Acoustical Society of America*, *137*(3), 1513-1528.

Dellwo, V., & Schmida, S. (2016). Speaker-individual rhythmic characteristics in read speech of German-Italian bilinguals. *TRENDS IN PHONETICS AND PHONOLOGY. STUDIES FROM GERMAN-SPEAKING EUROPE*, 349.

Duez, D. (1999). La fonction symbolique des pauses dans la parole de l'homme politique. *Faits de langues*, *7*(13), 91-97.

Eskénazi, M. (1993). Trends in Speaking Styles Research. In *Proceedings of the 3rd European Conference on Speech Communication and Technology* (p. 501-509). Berlin, Germany.

Fónagy, I. (1983). *La vive voix: essais de psycho-phonétique* (Vol. 20). Payot.

Labov, W. (1972). *Sociolinguistic patterns* (University of Pennsylvania Press). Philadelphia.

Leemann, A., Kolly, M.-J., & Dellwo, V. (2014). Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. *Forensic Science International*, *238*, 59-67.

Ramus, F. (1999). La discrimination des langues par la prosodie: Modélisation linguistique et études comportementales. *De la caractérisation…… à l'identification des langues*, 131.

Simon, A.-C., Auchlin, A., Avanzi, M., & Goldman, J.-P. (2010). Les phonostyles: une description prosodique des styles de parole en français. *Les voix des Français: en parlant, en écrivant, Bern: Lang*, 71–88.

Xu, Y. (2013). ProsodyPro—A tool for large-scale systematic prosody analysis. In *TASP'2013*. Aix en Provence, France: Laboratoire Parole et Langage, France.

# Variabilité inter et intra locuteurs de mesures spectrales et prosodiques en parole lue

Cédric Gendrot, Gabriele Chignoli, Nicolas Audibert, Cécile Fougeron
Laboratoire de Phonétique et Phonologie
19 rue des bernardins 75005 Paris
cedric.gendrot@univ-paris3.fr, gabriele.chignoli@gmail.com,
nicolas.audibert@univ-paris3.fr, cecile.fougeron@univ-paris3.fr

## RESUME

Cette étude préliminaire tente de déterminer des paramètres spectraux et prosodiques qui délimitent la variabilité inter et intra locuteurs pour un corpus de parole lue par 9 locuteurs d'une liste de mots et du texte « la bise et le soleil », enregistré tous les ans depuis 2012, incluant 3 répétitions à chaque enregistrement. Parmi l'ensemble des paramètres testés, nous montrons que s'avèrent pertinents le Centre de Gravité spectral (CoG), la f0 moyenne - pour les paramètres identifiés auparavant dans la littérature-, mais également le ratio harmoniques sur bruit (HNR), l'indice de variabilité temporelle entre les segments (Cnpvi/Vnpvi), ainsi que les variations syntagmatiques de f0.

## ABSTRACT

This preliminary study aims at distinguishing spectral and prosodic parameters that delimitate inter and intra speakers' variability for a corpus of words and a small text ("la bise et le soleil") read by 9 speakers every year since 2012, with 3 repeats for each recording. Amongst all tested parameters, we show that spectral Center of Gravity (CoG), mean f0 – for parameters mentioned previously in the literature-, but also the Harmonics to Noise Ratio (HNR), the index of temporal variability between segments (Cnpvi/Vnpvi), as well as syntagmatic variations of f0.

MOTS-CLES : variabilité inter locuteurs, variabilité intra locuteurs, moments spectraux, rythme, prosodie.

KEYWORDS: inter speakers' variability, intra speakers' variability, spectral moments, rhythm, prosody.

# 1 Etat de l'art

Les travaux en phonétique se concentrent majoritairement sur la recherche d'invariants au sein du signal acoustique pour un phénomène linguistique tel que par exemple l'accentuation et l'organisation


10.21437/JEP.2018-6

prosodique, et visant - au mieux - à la mise en évidence de quelques stratégies bien identifiées. Le signal acoustique de parole est soumis à de fortes variations, parmi lesquelles le contexte segmental, le contexte prosodique, les méthodes d'enregistrement du signal, le type de parole, etc. Dans le cadre de ce travail, nous considérons le locuteur comme un facteur de variation supplémentaire et notre but principal est de déterminer ses limites de variations, se rapprochant ainsi des travaux portant sur la reconnaissance du locuteur. Puisque le choix des extraits de parole peut impliquer une forte variation, nous avons choisi une analyse contrôlée avec un protocole d'enregistrement et un corpus identiques pour tous les locuteurs.

Pour les systèmes automatiques de reconnaissance du locuteur, l'objectif est de chercher les caractéristiques propres au locuteur, il est généralement admis que ces systèmes procèdent en trois phases : une phase de paramétrisation du signal acoustique, avant de déterminer un modèle du locuteur qui permettra au final d'aboutir à la phase de décision. Le taux d'erreurs ('Equal Error Rate' qui délimite le seuil entre le taux de fausses acceptations et de faux rejets) pour des extraits de 2,5 minutes avoisine les 5,3% pour les meilleures séries. Nous ne procéderons à aucune de ces phases, en nous concentrant sur les mesures acoustiques utilisées classiquement en phonétique expérimentale, et qui donnent des indices concrets sur la performance articulatoire ou le timbre du locuteur. Nous nous positionnons dans la continuité du travail de Kahn (2011) qui cherchait les indices acoustiques pertinents pour distinguer plusieurs locuteurs au moyen de la taille d'effet (éta carré ou $\eta^2$, Levine & Hullet, 2002) du facteur locuteur considéré comme variable dépendante lors d'une ANOVA. Les valeurs de F et de p ne sont pas présentées car elles s'avèrent systématiquement significatives et nous cherchons à mettre en valeur l'importance de l'effet, plus que sa significativité.

Dans les systèmes automatiques de reconnaissance du locuteur, l'information temporelle a été majoritairement abandonnée (Haton et al., 2006) ou alors considérée uniquement d'après les variations d'une trame d'analyse à une autre (en moyenne 15 à 20 ms), et nous choisirons ici d'inclure également des mesures prenant en compte la variation temporelle, depuis le niveau phonémique jusqu'à la totalité de la production. Si certains systèmes ont malgré tout tenté d'incorporer des informations temporelles comme la durée des mots, des phonèmes et des pauses (Shriberg et Stolcke, 2008), voire également en modélisant la courbe de fréquence fondamentale (Kockmann et al., 2010), ceux-ci étaient dans tous les cas combinés à des systèmes « classiques », i.e. intégrant des coefficients cepstraux par ailleurs.

Ce travail permettra une meilleure connaissance du facteur locuteur et de ses limites de variation dans un cadre expérimental contrôlé. L'objectif de ce travail pourrait avoir pour but d'améliorer les connaissances en matière de reconnaissance du locuteur, mais également de mieux identifier les invariants dans le signal acoustique. Nous n'aborderons pas ici la capacité humaine à identifier un locuteur par sa voix. Kahn (2011) a montré que les performances des auditeurs dans ce cadre sont excessivement dépendantes des conditions et de la tâche demandée (longueur des extraits présentés, connaissance préalable des locuteurs, état émotionnel des locuteurs, etc.), ainsi que de leur capacités.

Nous cherchons les indices idiosyncratiques des locuteurs contenus dans le signal acoustique. Comme décrit par Kahn (2011), ces indices ne sont pas uniformément répartis dans le signal de parole, ce qui implique que la pertinence des indices acoustiques varie en fonction des extraits de parole. De par le choix d'un corpus lu et identique pour tous les locuteurs, les résultats présentés ici sont considérés comme 'text dependent', mais nous aborderons dans la discussion l'apport que ces résultats peuvent fournir sur des systèmes de reconnaissance du locuteur dits 'text-independent'.

# 2 Protocole expérimental

## 2.1 Corpus

Les données acoustiques analysées ici correspondent à des extraits du corpus PATATRA (Parole AdulTe A TRavers les Ages, élaboré et recueilli au Laboratoire de Phonétique et Phonologie) (Fougeron et al., 2017), qui consiste en un enregistrement annuel de 9 locuteurs (4 hommes et 5 femmes) depuis 2012, soit actuellement 5 années, incluant une liste de mots, un texte lu (« la bise et le soleil »), des exercices de phonation, et une courte séquence de parole spontanée. Seuls la liste de mots et le texte sont utilisés pour l'analyse des données dans le présent travail. Ces items étant systématiquement répétés trois fois pour chaque enregistrement, nous avons donc au total 5 années * 3 répétitions * 9 locuteurs pour une production de 58 mots et un texte. Les locuteurs sont enregistrés au moyen d'un microphone casque AKG C 520, après calibration avec un sonomètre. Ce corpus a été conçu pour évaluer le vieillissement de la voix, mais nous l'exploitons ici pour au contraire valider la cohérence des mesures acoustiques d'une année à la suivante, tout en multipliant les répétitions.

## 2.2 Mesures acoustiques

Dans les travaux de Kahn (2011), les mesures acoustiques effectuées comprennent la f0, le jitter et le shimmer mesurés sur les voyelles, les formants 1 à 4 mesurés sur les voyelles orales, et pour l'ensemble des phonèmes le centre de gravité spectral, la durée, ainsi que les MFCC (Mel Frequency Cepstral Coefficients) et les LFCC (Linear Frequency Cepstral Coefficients).

Dans ses travaux, le centre de gravité spectral a été mesuré comme un paramètre fiable de variation chez le locuteur, notamment pour les fricatives et les nasales (en comparaison des occlusives). Pour les voyelles orales, plus la voyelle est ouverte et antérieure, plus l'effet du locuteur est élevé sur les valeurs de centre de gravité. Ce dernier résultat se confirme sur les valeurs de formants (et notamment F3 et F4). Les transitions formantiques (Mc Dougall, 2006) s'avèrent moins pertinentes. Il apparaît en résumé que les voyelles ne discriminent pas de façon égale les différents locuteurs, les voyelles ouvertes et les voyelles nasales étant les plus informatives. L'hypothèse avancée pour les voyelles nasales est que l'intégration de la cavité nasale par son ouverture fournit une information supplémentaire par la qualité des résonances qui sont propres aux locuteurs. Il en serait de même pour les voyelles ouvertes pour lesquelles le conduit vocal est plus large.

Nous nous proposons de prolonger ces travaux en partant des mesures de CoG afin de comparer nos résultats à ceux de Kahn (2011) dans un premier temps, puis d'étendre nos mesures aux autres moments spectraux ('skewness', 'standard deviation', 'kurtosis'), mais également des mesures de rapport Harmoniques sur bruit (HNR) qui déterminent la proportion de bruit et de voisement dans le signal acoustique, les variations mélodiques (minimum, maximum, étendue et pente de f0) sur les mots et l'énoncé. Toutes les mesures ont été effectuées à l'aide de PRAAT en utilisant les paramètres par défaut, la segmentation en phonèmes a été effectuée par un aligneur (EasyAlign), puis corrigée manuellement par les 2 premiers auteurs. La combinaison de l'effet de ces différents paramètres sera également évaluée puisqu'ils pourraient s'avérer complémentaires dans la distinction entre différents locuteurs. Les statistiques présentées dans les sections suivantes ont été effectuées avec R (version 3.4.2. 2017).

# 3 Analyses

## 3.1 Analyses sur les mots

Une ANOVA est effectuée pour chaque phonème avec comme variable dépendante chaque valeur acoustique (voir 2.2.), et comme variable indépendante (facteur fixe) les locuteurs et l'année d'enregistrement, afin de mesurer l'influence des locuteurs sur les mesures acoustiques. Nous ne présentons pas ici les résultats pour les 3 répétitions par année car leur taille d'effet est systématiquement proche de 0. Les résultats présentés dans les tables 1 et 2 ci-dessous confirment les résultats observés par Kahn (2011), à savoir que les fricatives, les nasales, mais également les sonantes présentent une taille d'effet plus importante que les occlusives, de même les voyelles antérieures, ouvertes et nasales, comparativement aux voyelles fermées et postérieures. Les paramètres de CoG, auquel nous ajoutons le HNR et le maximum de f0 mesurée sur le mot sont donc des facteurs pertinents pour distinguer les locuteurs, notamment pour certains phonèmes comme /m/, /l/ ou /ɑ̃/ Le facteur « année » (d'enregistrement) présente des valeurs de taille d'effet entre 0 et 0.06, ce qui montre que ce facteur n'est pas sensible à nos mesures, il sera donc éludé pour la suite de cet article.

Les autres mesures acoustiques effectuées, à savoir les trois autres moments spectraux ('kurtosis', 'standard deviation' et 'skewness' révèlent des valeurs semblables mais inférieures à celles présentées pour le CoG et ne sont donc pas présentées ici ; de même pour les autres mesures de f0 (moyenne, étendue, minimum et pente mesurés sur chaque mot).

| Taille d'effet | | /b/ | /d/ | /p/ | /t/ | /k/ | /f/ | /v/ | /ʃ/ | /ʒ/ | /l/ | /m/ | /ʁ/ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| COG | Année | 0,00 | 0,00 | 0,11 | 0,00 | 0,02 | 0,02 | 0,06 | 0,00 | 0,00 | 0,01 | 0,01 | 0,02 |
| | Locuteur | 0,15 | 0,27 | 0,13 | 0,04 | 0,07 | **0,43** | 0,29 | **0,42** | 0,40 | **0,45** | **0,47** | 0,32 |
| HNR | Année | 0,01 | 0,01 | 0,01 | 0,02 | 0,01 | 0,01 | 0,02 | 0,02 | 0,01 | 0,00 | 0,02 | 0,03 |
| | Locuteur | 0,33 | 0,31 | 0,07 | 0,12 | 0,03 | **0,13** | **0,28** | **0,33** | **0,26** | **0,33** | **0,49** | 0,12 |
| maximum f0 | Année | 0,00 | 0,00 | | | | | 0,03 | | 0,00 | 0,01 | 0,00 | 0,00 |
| | Locuteur | 0,83 | 0,80 | | | | | **0,62** | | **0,56** | **0,60** | **0,92** | 0,29 |

TABLE 1 : Tailles d'effet du facteur locuteur mesuré sur chaque consonne pour les mesures de CoG, HNR et f0

| Taille d'effet | | i | u | a | ɑ̃ |
|---|---|---|---|---|---|
| COG | Année | 0,00 | 0,00 | 0,03 | 0,01 |
| | Locuteur | **0,47** | **0,40** | **0,55** | **0,62** |
| HNR | Année | 0,01 | 0,01 | 0,01 | 0,02 |
| | Locuteur | **0,33** | **0,31** | 0,07 | 0,12 |
| maximum f0 | Année | 0,00 | 0,00 | 0,00 | 0,04 |
| | Locuteur | **0,83** | **0,80** | **0,64** | **0,96** |

TABLE 2 : Tailles d'effet du facteur locuteur mesuré sur chaque voyelle pour les mesures de CoG, HNR et f0

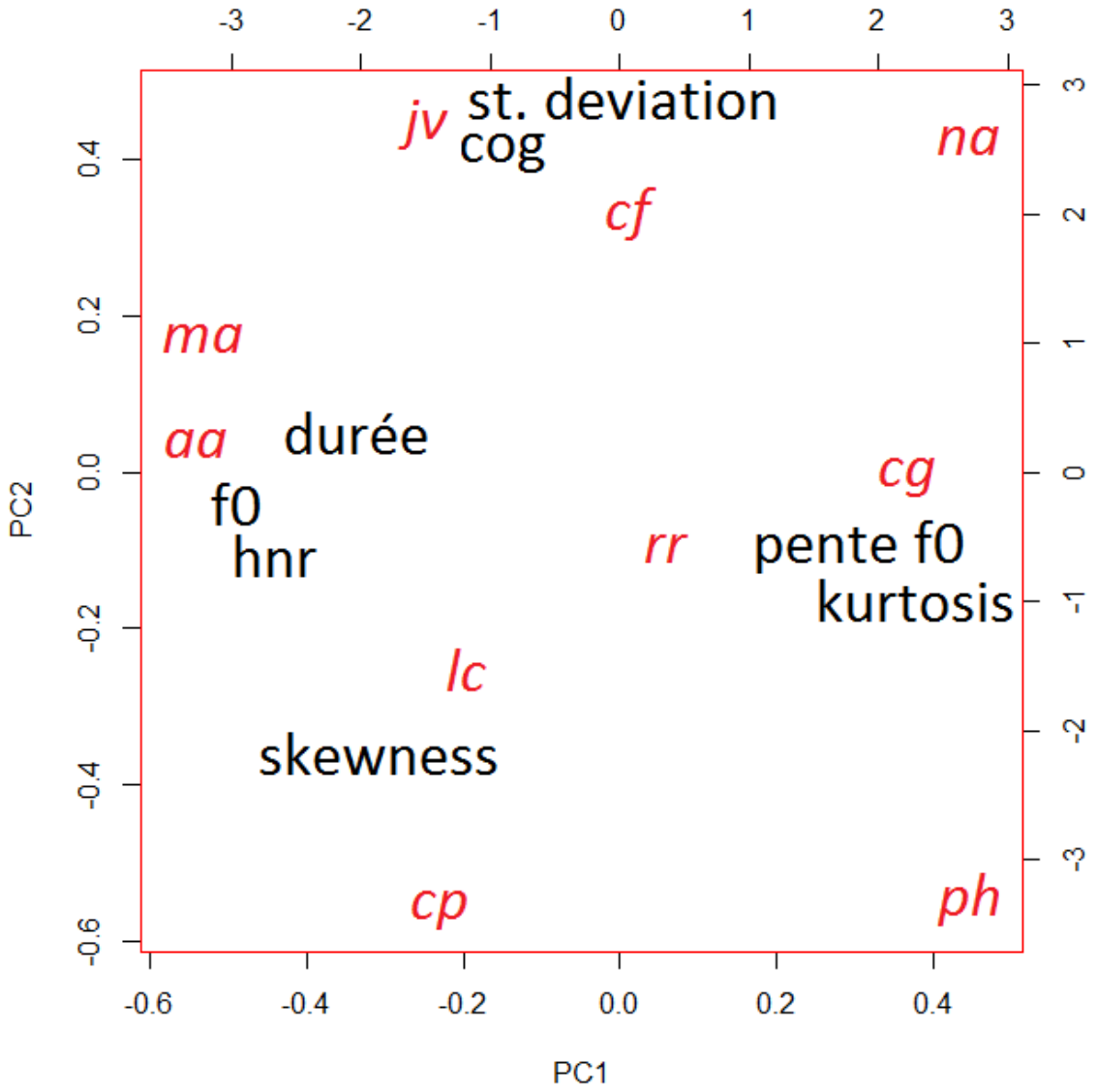



FIGURE 1: Analyse en Composantes Principales des mesures acoustiques effectuées sur /ɑ̃/

Nous présentons maintenant les analyses qui combinent les différents facteurs mesurés afin de tester leur complémentarité. Les analyses ont été effectuées sur / ɑ̃ / car c'est le phonème ayant montré la taille d'effet la plus élevée sur les tables 1 et 2 pour le CoG et le maximum de f0 (seul le HNR est plus pertinent pour les consonnes.

Une Analyse en Composantes Principales (figure 1), avec les valeurs de chaque locuteur moyennées par année et répétition, montre que sont orthogonaux les mesures de CoG et de HNR, alors que les autres paramètres leurs sont corrélés. La proportion de variance de la 3ème composante étant de seulement 8.5% (comparativement à 60% et 21% pour les 2 premières), nous ne présentons ici que les 2 premières composantes principales. Par la suite, nous avons effectué un arbre de classification pour comprendre dans quelle mesure ces paramètres retenus interagissent entre eux pour classifier les

locuteurs en fonction de leur variance. La figure 2 montre que les paramètres de f0, de CoG et de HNR interagissent pour délimiter les différents locuteurs et doivent donc être pris en compte simultanément et non un par un.

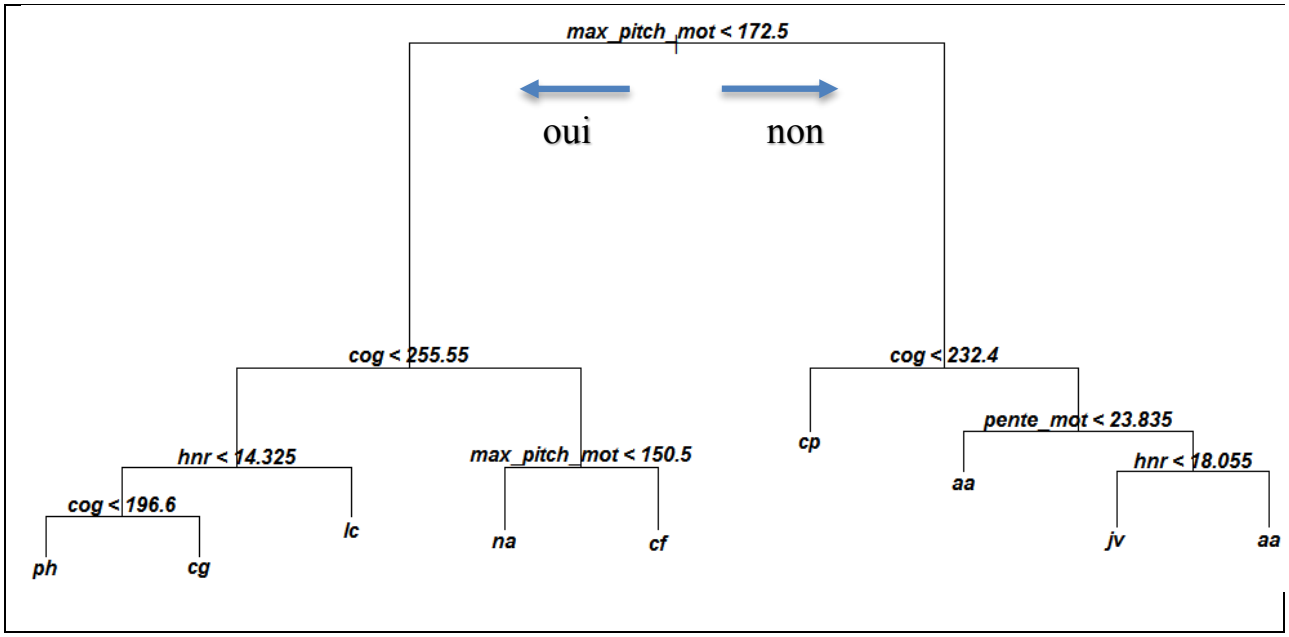



FIGURE 2: Arbre de classification : interaction entre les différentes mesures de CoG et les mesures de HNR effectuées sur la voyelle /ɑ̃/

## 3.2 Analyses sur le texte « La bise et le soleil »

Dans cette section, nous présentons les résultats des analyses effectuées sur le texte « la bise et le soleil ». Seules les 5 dernières séquences « alors le soleil a commencé à briller », « et au bout d'un moment », « le voyageur réchauffé a ôté son manteau », « ainsi la bise a dû reconnaître » et « que le soleil était le plus fort des deux » ont été analysées, les autres étant actuellement en cours de correction manuelle. Les mesures de rythme que nous proposons ici ont été calculées à l'aide de Correlatore 2.2 (Mairano & Romano, 2010) qui permet de calculer %V, ΔV, ΔC, VarcoV, VarcoC, rPVI, nPVI et CCI définis ci-dessous.

– %V : Pourcentage de la durée de la phrase composée de ses intervalles vocaliques
– ΔV/C : Ecart-type des intervalles de durée vocaliques / consonantiques
– VarcoV/C : Ecart-type des intervalles de durée vocaliques / consonantiques divisé par la durée moyenne des intervalles vocaliques / consonantiques (et multiplié par 100)
– CnPVI / VnPVI : Consonant / Vocalic Normalized pairwise variability index; moyenne des différences entre consonnes / voyelles successives, divisées par la moyenne des deux durées)
– CCI : Control/Compensation Index. Normalisation du rPVI; moyenne des différences entre intervalles successifs, où chaque intervalle est divisé par le nombre de segments dans l'intervalle.

Après application de l'ensemble de ces mesures, les variations intra-locuteurs et les différences inter-locuteurs les plus importantes ont été observées pour les mesures de CnPVI et VnPVI que nous présentons dans la figure 3.

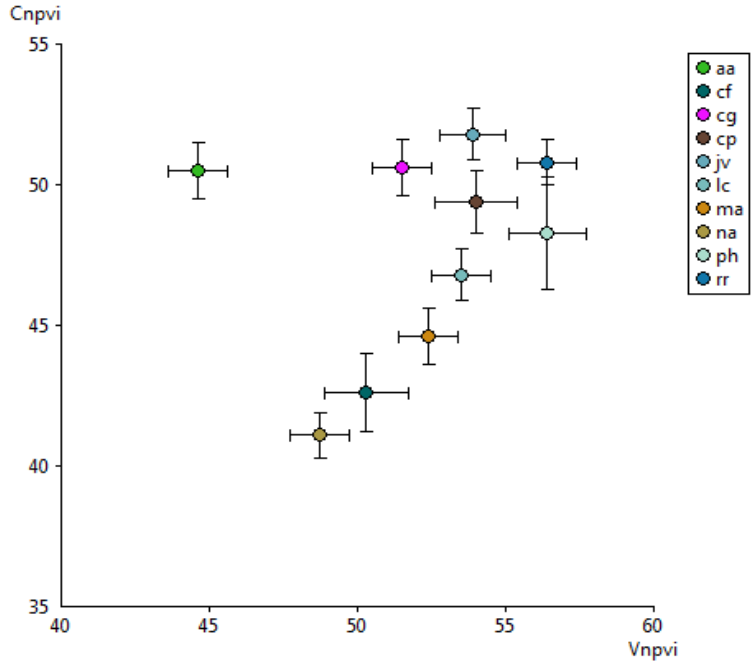



FIGURE 3 : VnPVI sur CnPVI pour le texte « la bise et le soleil »

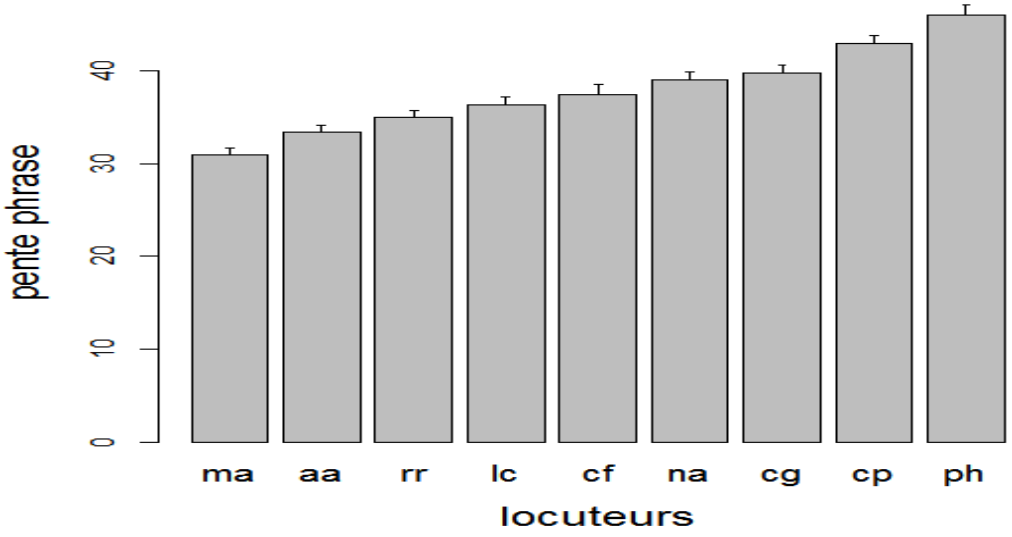



FIGURE 4 : Mesure de pente de f0 (en Hz/s) sur les phrases du texte « la bise et le soleil »

La pente de f0 mesurée comme la différence en demi-tons sur toute la longueur de la phrase (valeur absolue moyenne). La taille d'effet du locuteur sur la pente de f0 mesurée pour une ANOVA avec

comme variables dépendantes la pente de f0 et comme facteurs fixes le locuteur et l'année d'enregistrement est de 0.56. Nous voyons sur la figure 4 comment les locuteurs se répartissent sur les différentes valeurs de pente.

# 4 Discussion et conclusion

Nous avons montré dans cette étude préliminaire que d'autres paramètres s'avéraient pertinents et complémentaires de ceux déjà connus dans la littérature (valeurs moyenne de f0, CoG, formants), à savoir le HNR (ratio harmoniques sur bruit), la pente de f0 mesurée sur le mot ou la phrase et le rythme (VnPVI/CnPVI) dans l'énoncé, ainsi que la combinaison de ces paramètres. Il sera bien sûr nécessaire de tester ces paramètres sur un nombre plus important de locuteurs. Nous avons également montré que les autres moments spectraux ('skewness', 'kurtosis' et 'standard deviation') n'apportent qu'une information redondante par rapport au centre de gravité spectral. Comme observé par Kahn (2011), les segments qui montrent une taille d'effet plus conséquente dans notre corpus sont les nasales (consonnes et voyelles), les sonantes, et dans une moindre mesure les fricatives. Il semble que les segments porteurs d'informations idiosyncratiques relèvent souvent des différences morphologiques, comme par exemple la forme des fosses nasales pour les voyelles nasales, la forme des dents et du palais pour les fricatives dentales et palatales, etc. En effet, les caractéristiques propres au locuteur consistent en deux éléments principaux : les variations physiques statiques dans le signal qui correspondent aux caractéristiques physiques des articulateurs mises en évidence ici par le CoG, le HNR et la valeur moyenne de f0, alors que les variations dynamiques reflètent les différences comportementales qui seraient plus vraisemblablement révélées par le rythme et la pente de f0.

Kahn (2011) soulignait en conclusion de sa thèse la nécessité de distinguer le locuteur et l'extrait de parole dans une tâche d'identification des indices idiosyncratiques du locuteur. Elle mentionnait également que plus la parole est contrôlée, plus le locuteur est contraint dans son énonciation, plus le risque de ne pas trouver d'indices discriminants pour le locuteur est grand. Le corpus PATATRA comprend également des séquences de parole spontanée pour chaque enregistrement que nous analyserons dans la suite de ce travail, le but premier ayant été ici de déterminer les paramètres acoustiques les plus pertinents et leur complémentarité dans des contextes phonémiques contrôlés. Les mesures présentées ici sont pour la plupart des mesures dites 'text-dependent' puisqu'elles impliquent un étiquetage et un alignement en phonèmes préalable, mais les mesures liées à la f0 (maximum et pente) pourraient être effectuées sans connaissances linguistiques à priori. Dans la suite de ce travail, nous viserons également à tester si les mesurées classiquement effectuées par les systèmes automatiques de reconnaissance du locuteur, à savoir les MFCC/LFCC (MEL Frequency /Linear Predictive Cepstral Coefficients), montrent un effet du locuteur plus important que les paramètres mesurés ici.

# Remerciements

Nous remercions l'ANR VOXCRIM (ANR-17-CE39-0016) ainsi que le LaBeX Empirical Foundations of Linguistics (EFL).

# Références


FOUGERON, C., DELVAUX, V., GENDROT, C., LAGANARO, M., MENARD, L. (2016). Introducing two databases of spoken French throughout adulthood. in Workshop on Speech Perception and Production across the Lifespan, London, England, 2017

HATON, J., CERISARA, C., FOHR, D., LAPRIE, Y. ET SMAÏLI, K. (2006). Reconnaissance automatique de la parole, Du signal à son interprétation. Paris : Dunod.

KAHN, J. (2011). Parole de locuteur : performance et confiance en identification biométrique vocale, Thèse de Doctorat, Avignon.

KOCKMANN, M., BURGET, L. ET CERNOCKY, J. (2010). Investigations into prosodic syllabe contour features for speaker recognition. in International Conference in Acoustics, Speech and Signal Processing (ICASSP), Dallas, 4418–4421.

LEVINE, T. ET HULLET, C. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. Human Communication Research 28(4), 612–625.

MAIRANO, P. ET ROMANO, A. (2010). Un confronto tra diverse metriche ritmiche usando Correlatore. In: Schmid, S., Schwarzenbach, M. & Studer, D. (eds.) La dimensione temporale del parlato, Proc. of the V Natioanl AISV Congress (Associazione Italiana di Scienze della Voce) (University of Zurich, Collegiengebaude, 4th-6th February 2009), Torriana (RN): EDK, 79-100.

MCDOUGALL, K. (2006). Dynamic features of speech and characterization of speakers : towards a new approach using formant frequencies. Speech, Language and the Law 13, 89–126.

R CORE TEAM (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

SHRIBERG, E. ET STOLCKE, A. (2008) The case for automatic Higher-Level features in forensic speaker recognition. Dans les actes de International Conference on Speech Communication and Technology (Interspeech), Brisbane, 1509–1512.

# Le /R/ « roulé » en français et dans quelques langues régionales de France

Timothée Premat[1,2] & Philippe Boula de Mareüil[2]
(1) SFL (UMR 7023), 59/61, rue Pouchet 75017 Paris, France
(2) LIMSI, CNRS & Univ. Paris-Saclay (UPR 3251), 508, rue John von Neumann, 91405 Orsay, France
timothee.premat@etud.univ-paris8.fr, mareuil@limsi.fr

RÉSUMÉ
Ce travail vise à documenter la prononciation du /R/ chez des locuteurs (souvent âgés) qui, en plus du français, parlent un dialecte gallo-roman et qui « roulent » les *r*, de façon plus ou moins stable, en français et dans leur dialecte : gascon, languedocien, francoprovençal, bourguignon et mainiot. Outre qu'on trouve davantage de [r] apicaux (« roulés ») en dialecte qu'en français, il ressort de ce travail que les [r] sont davantage présents en attaque de syllabe ou en coda interne qu'en fin de mot, et chez les locuteurs âgés davantage que chez les plus jeunes – on ne trouve pas de [r] apical en français régional chez nos locuteurs de moins de 65 ans. Face à cette situation contrastée, nous avançons quelques éléments de discussion à propos de l'interface entre phonologie et sociolinguistique.

ABSTRACT
**Variation of the rhotic phoneme in French and regional languages of France**

This work aims at documenting the pronunciation of the /R/ rhotic in (often elderly) speakers who, in addition to French, speak a Gallo-Roman dialect and who produce an apical ("rolled") *r*, in a more or less stable fashion, in French and their dialect : Gascon, Languedocien, Francoprovençal, Bourguignon or Mainiot. It turns out that there are more apical [r]s in dialects than in French, more so in a syllable onset or an internal coda than in a word-final position, and more so in elderly speakers than in younger ones. By contrast, we did not find any apical [r] in the regional French varieties of speakers under 65 years old. Faced with this situation, we offer some elements of discussion about the interface between phonology and sociolinguistics.

MOTS-CLÉS : Variation, standardisation, dialectologie, phonème /R/, Gallo-Romania.

KEYWORDS: Variation, standardisation, dialectology, /R/ phoneme, Gallo-Romania.

# 1 Introduction

L'archiphonème rhotique /R/, dont la prononciation peut prendre des formes multiples (car phonologiquement sous-spécifié), est à l'origine de travaux pionniers en sociolinguistique (Labov, 1972). Le [r] « roulé », antérieur et apical (prononcé avec la pointe de la langue) et hérité du *r* latin, correspondait encore au bon usage pour le français du XVII$^{e}$ siècle (Rouillé, 2008, p. 11). C'était celui qu'enseignait le Maître de philosophie à Monsieur Jourdain, dans *Le Bourgeois gentilhomme* de Molière (1673). Il a, depuis, largement été remplacé par un [ʁ] postérieur, dorsal (impliquant le dos de la langue), mais la vitesse et les conditions de ce changement phonétique varient selon les régions.

L'origine parisienne du passage de [r] à [ʁ] est assez consensuelle, même si son caractère progressif ou spontané reste disputé. L'articulation postérieure serait apparue, entre le XVII$^{e}$ et le XVIII$^{e}$ siècle,


10.21437/JEP.2018-7

dans le basilecte parisien, avant d'être rapidement adoptée par l'acrolecte standard. Jusqu'à une date relativement récente, toutefois, cette innovation n'a pas réussi à s'imposer sur l'ensemble du territoire, et le [r] apical a pu se maintenir davantage dans les dialectes (ou langues régionales) traditionnels. Cela tient au fait que la standardisation du français, dans le monde rural, ne s'est achevée qu'il y a peu. On peut estimer, à partir d'un faisceau de sources concordantes, que c'est seulement depuis le début du XX$^{e}$ siècle que le français standard a été compris sur l'ensemble du territoire, et que c'est uniquement depuis les années 1960 que l'usage réel des dialectes est devenu marginal (Lodge, 1997, pp. 269–272). Cette situation diverge selon les aires linguistiques, et a concerné les dialectes occitans et francoprovençaux plus tardivement que les dialectes d'oïl (Lodge, 1997, pp. 253–254).

L'*Atlas Linguistique de la France* (Gilliéron & Edmont, 1902–1910) montre que les rhotiques postérieures sont largement absentes des dialectes français ruraux à la fin du XIX$^{e}$ siècle, y compris dans les dialectes proches de Paris. Les atlas ultérieurs pour la plupart, à l'exception de ceux du Languedoc (Ravier *et al.*, 1978 ; Boisgontier *et al.*, 1981), n'enregistrent pas systématiquement les différents allophones du /R/, arguant pour certains que cette variation n'est pas de caractère diatopique (Gardette, 1990 ; Guillaume *et al.*, 1975 ; Martin & Tuaillon, 1978 ; Taverdet, 1975). Il convient donc d'exploiter de nouvelles données pour mettre au jour certains facteurs de cette variation.

Dans le cadre de l'*Atlas sonore des langues régionales de France* (Boula de Mareüil *et al.*, 2018), des enquêtes de terrain ont été menées en 2014–2017, entre autres auprès de locuteurs d'occitan (gascon ou languedocien), de francoprovençal, ainsi que dans le domaine d'oïl, auprès de locuteurs de bourguignon et de mainiot (Sarthe). Plus de 200 locuteurs ont été enregistrés, représentant environ 200 heures de parole (lue et spontanée). Les locuteurs ont notamment été enregistrés lisant un même texte en français et le traduisant en langue régionale. Parmi eux, certains produisent des [r] apicaux dans leur langue régionale et parfois en français. Nous n'avons pas relevé de [r] apicaux dans des régions comme la Normandie, la Picardie et la Lorraine. Nous en avons relevé au Pays basque, en Bretagne, en Corse, en Roussillon (catalan) et en Alsace, mais nous n'avons pas retenu ces régions dans la présente étude, qui se concentre sur les parlers gallo-romans *stricto sensu*.

Quels sont les contextes phonologiques et les facteurs sociaux favorisant l'apparition de ce [r] ? C'est cette double question que nous nous proposons d'étudier ici. Après une présentation du corpus et de la méthode employés (section 2), nous analyserons les facteurs phonologiques liés à la position dans la syllabe (section 3) et les facteurs sociolinguistiques liés notamment à l'âge des locuteurs (section 4), avant de conclure cet état des lieux (section 5).

# 2 Corpus et méthode

Le corpus analysé s'appuie sur la lecture de la fable d'Ésope « La bise et le soleil », utilisée depuis plus d'un siècle par l'Association Phonétique Internationale pour décrire un grand nombre de langues et dialectes du monde. Ce texte a été lu en français (120 mots, correspondant à une minute de parole) et traduit par les enquêtés : en gascon, en languedocien, en francoprovençal, en bourguignon et en mainiot. Dans chacun de ces 5 dialectes, nous avons sélectionné 7 locuteurs possédant des [r] apicaux dans leurs systèmes phonétiques. Ces locuteurs (28 hommes et 7 femmes), sont nés entre 1930 et 1980. Plutôt engagés sur le terrain culturel et linguistique, ils sont issus de milieux socioprofessionnels variés (universitaire, artiste, enseignant, cheminot, postier, gendarme, paysan ou bûcheron, par exemple). La figure 1 donne leur âge minimal, maximal et moyen par dialecte : on peut constater que l'âge moyen de nos locuteurs d'oïl et de francoprovençal est plus élevé (74 ans) que celui de nos locuteurs occitans (62 ans), en raison notamment de la présence de locuteurs de moins de 40 ans dans notre échantillon méridional. La moyenne globale est de 70 ans. Le faible nombre

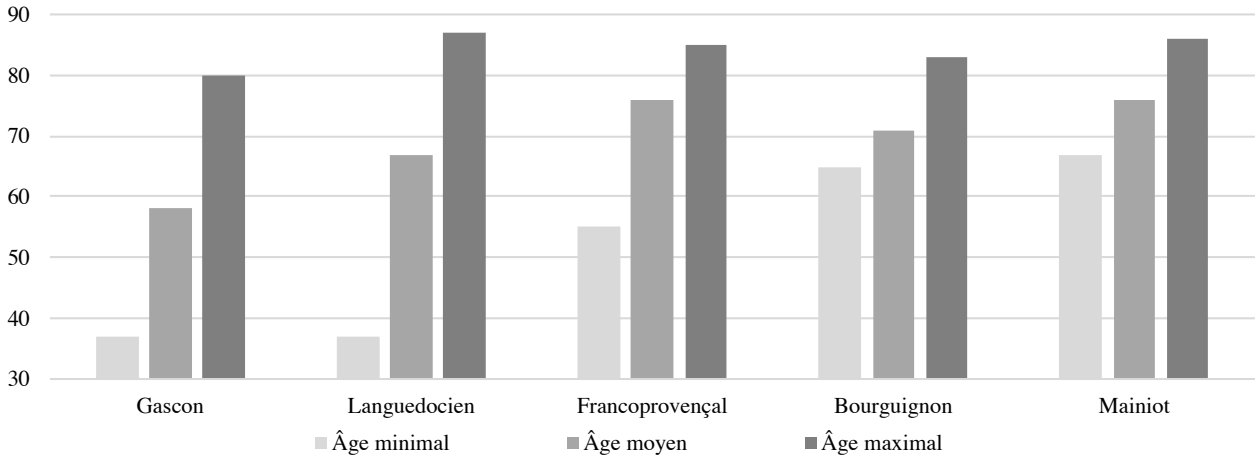



FIGURE 1 – Âge (min., max. et moyen) des locuteurs des différents dialectes

de locutrices – par ailleurs explicable en termes sociolinguistiques (Labov, 1972) – ne permet pas d'interpréter de façon genrée la réalisation des /R/ de notre échantillon.

Pour chaque locuteur, nous avons mesuré les pourcentages de [r] apicaux en français et en dialecte, permettant ainsi des comparaisons, toutes choses égales par ailleurs. Chacune des 70 versions de la fable retenues totalise entre 20 et 25 occurrences de /R/. Sans compter les cas où le phonème n'est pas réalisé, ce sont 1689 /R/ qui ont été analysés.

Il faut en premier lieu noter que presque tous les allophones possibles du /R/ semblent présents sur le territoire gallo-roman. Au-delà des [r], [ʀ] et [ʁ] (et de sa variante contextuelle [χ]), on trouve dans nos données la monovibrante alvéolaire [ɾ], l'approximante alvéolaire [ɹ] et la rétroflexe [ɻ]. On trouve également des cas où le /R/ n'est pas produit sous forme rhotique : il peut se transformer en consonne non-rhotique – notamment [l] mais aussi [ð, θ, z] (Straka, 1979), en voyelle ou en approximante médiane non-rhotique (p. ex. FORTIS → *foa* [fwa] en francoprovençal), être assimilé à l'approximante centrale ou latérale qui suit (p. ex. *arribariá* [ariβajɔ] en languedocien et *alors la (bise)* [alɔlːa] en bourguignon), voire n'être pas produit du tout. La disparition de la rhotique, en particulier, est courante lorsque /R/ est branché en attaque avec une occlusive, pour reprendre une analyse phonologique désormais classique (Dell, 1976) ; mais le signal acoustique peut garder des traces de sa présence, qui demeure identifiable perceptivement (Gendrot, 2014).

Dans le travail rapporté ici, nous avons uniquement distingué les réalisations rhotiques selon leur lieu d'articulation, dans une typologie binaire : rhotique antérieure (de loin le plus souvent l'apicale [r]) *vs* postérieure (majoritairement [ʁ]). L'annotation a été faite par le premier auteur de cet article, sur la base de la perception et, le cas échéant, de l'inspection des spectrogrammes. Dans quelques cas délicats (Engstrand *et al.*, 2007), le second auteur est intervenu et un accord a été trouvé. Nous nous sommes également intéressés à la position en attaque de syllabe ou en coda, cette dernière (en finale ou devant consonne) ayant été décrite par Morin (2013), parmi d'autres, comme favorisant l'affaiblissement du [r] en [ʁ]. Nos données montrent cependant que ce conditionnement n'est pas suffisant : contrairement à ce qui a pu être décrit par exemple pour le français de Belgique (Demolin, 1999), nous ne trouvons pas dans notre corpus de distribution phonologique univoque des allophones rhotiques. Certains locuteurs, indépendamment du contexte phonologique, produisent même des séquences où toutes les rhotiques sont antérieures, puis d'autres où toutes les rhotiques sont postérieures. Précisons que, dans l'annotation, les resyllabations dues à des enchaînements ont été prises en compte (ex. *faire‿ôter*, où le /R/ de *faire* n'est pas annoté final mais antévocalique).

TABLE 1 – Pourcentage de rhotiques apicales et nombre d'occurrences de /R/ sur lequel porte ce pourcentage (en gris, entre parenthèses) en fonction du contexte : #__ indique une position initiale, V__V une position intervocalique, T__ une attaque branchante interne, T__# une attaque branchante en finale absolue, __# une position de finale absolue et __C une position de coda interne.

| | | Attaque | | | | Coda | | Somme |
|---|---|---|---|---|---|---|---|---|
| | | #__ | V__V | T__ | T__# | __# | __C | |
| Gascogne | dial. | 100 (7) | 84 (116) | 92 (12) | – (0) | 100 (9) | 96 (28) | 88 (172) |
| | fr. | 14 (28) | 14 (44) | 14 (14) | 29 (7) | 15 (82) | 14 (14) | 15 (189) |
| Languedoc | dial. | 67 (21) | 85 (92) | 100 (16) | – (0) | 87 (15) | 70 (44) | 81 (188) |
| | fr. | 14 (28) | 15 (48) | 13 (15) | 29 (7) | 12 (77) | 7 (14) | 13 (189) |
| Aire fpr. | dial. | 46 (26) | 68 (31) | 78 (18) | 80 (5) | 32 (41) | 67 (6) | 54 (127) |
| | fr. | 36 (28) | 36 (42) | 42 (19) | 67 (3) | 11 (75) | 38 (13) | 27 (180) |
| Bourgogne | dial. | 86 (22) | 88 (34) | 94 (31) | 100 (1) | 90 (40) | 65 (20) | 86 (148) |
| | fr. | 43 (28) | 51 (43) | 47 (19) | 0 (3) | 41 (78) | 46 (13) | 44 (184) |
| Maine | dial. | 93 (14) | 85 (27) | 83 (30) | – (0) | 71 (34) | 67 (21) | 79 (126) |
| | fr. | 40 (25) | 27 (52) | 33 (15) | 25 (4) | 32 (75) | 33 (15) | 32 (186) |
| **Somme** | dial. | 72 (90) | 83 (300) | 89 (107) | 83 (6) | 68 (139) | 75 (119) | 79 (761) |
| | fr. | 29 (137) | 28 (229) | 32 (82) | 29 (24) | 22 (387) | 28 (69) | 26 (928) |

# 3 Résultats en fonction du contexte phonologique

Les résultats en fonction de l'environnement phonologique sont consignés dans la table 1, qui présente le pourcentage de réalisations apicales (et le nombre d'occurrences sur lequel porte ce pourcentage), pour chaque variété linguistique étudiée.

## 3.1 En français

Dans les variétés de français de Gascogne, du Languedoc, de l'aire francoprovençale (Pays de Savoie et Bresse), de Bourgogne et du Maine, les /R/ dorsaux sont largement majoritaires : ils concernent 74 % des réalisations. Dans notre échantillon, c'est dans l'aire occitane que les /R/ dorsaux sont les plus stables : 86 % des rhotiques y sont postérieures — voir §4 pour une explication en termes autres que géolinguistiques. Dans l'aire francoprovençale, les /R/ dorsaux sont moins stables (73 %) mais encore proches de la moyenne de nos 35 locuteurs. En zone d'oïl, la situation est contrastée : les /R/ dorsaux sont encore majoritaires dans le Maine (68 %) mais seulement légèrement (56 %) en Bourgogne, où nos locuteurs ont conservé de nombreux « *r* bourguignons ».

Parmi les contextes que nous avons relevés, seule la position de finale simple (à l'exclusion, donc, des attaques branchantes finales) semble avoir une influence décisive sur le lieu d'articulation de la rhotique dans plusieurs variétés (français de Bourgogne, de Bresse et des Pays de Savoie). Dans ce contexte, la variante postérieure [ʁ] est favorisée : 78 % des /R/ __# sont postérieurs, contre 68–72 % pour les autres positions. Cette tendance se retrouve, dans toutes les variétés, sous la forme d'un amuïssement courant de /R/ final (non relevé dans la table 1). On peut donc avancer l'analyse suivante : en position finale de mot, le /R/ est susceptible de s'affaiblir jusqu'à disparaître, comme c'est le cas du *-r* de l'infinitif des verbes du 1$^{er}$ groupe. Partant, dans un système contenant un [r] antérieur, l'allophone [ʁ] est en quelque sorte la version *faible* du /R/, située entre [r] et Ø.

L'autre position susceptible de favoriser cet affaiblissement, la coda interne, ne montre ce comporte-

ment qu'en français du Languedoc ; dans les autres dialectes, les pourcentages de réalisations apicales sont proches de la moyenne ou supérieurs à celle-ci. En français du Maine, la position initiale de mot (#__) montre un pourcentage de /R/ dorsaux (60 %) sensiblement inférieur à la moyenne (70 %).

## 3.2 En dialectes et langues minoritaires

Dans tous les dialectes des locuteurs sélectionnés, c'est le [r] apical qui l'emporte, même si celui-ci n'est pas hégémonique : il ne concerne que 79 % des rhotiques et, dans le cas du francoprovençal, n'est qu'à peine majoritaire (avec 54 % de production en zone alvéolaire).

Les dialectes occitans affichent une affinité très marquée pour les [r] apicaux. En gascon, cette affinité est plus forte en frontière de mot (100 % de [r] en initiale et finale absolues), en coda interne (96 %) et en attaque branchante (92 %) qu'à l'intervocalique (84 %), déjouant les prédictions d'affaiblissement en [ʁ] dans les contextes de coda. En languedocien, on observe des taux plus faibles de [r] à l'initiale (67 %) et en coda interne (70 %), tandis que dans toutes les autres positions ces taux sont au-dessus de 85 % et que toutes les rhotiques des attaques branchantes sont antérieures. Le traitement spécifique des attaques et des codas (internes et finales) du gascon peut s'expliquer par la conservation, chez certains locuteurs, d'une distinction étymologique entre rhotiques longues [rː] et brèves [ɾ], laquelle a pu évoluer en une opposition entre rhotiques postérieures et antérieures : [rː]→[ʁ] tandis que [ɾ] → [ɾ] (ou [r]) (Oliviéri & Sauzet, 2016). L'allongement des rhotiques initiales et la prothèse qui s'ensuit ([rːat] → [arːat] ou [aʁat]) expliquent le faible nombre d'occurrences initiales, tandis que les quelques occurrences non-prothétiques interviennent chez deux locuteurs n'ayant aucun [ʁ] postérieur dans leur idiolecte et chez un locuteur appliquant la distinction /rː/ ~ /ɾ/ sous forme [ʁ] ~ [r, ɾ], mais comme ce locuteur ne pratique ni l'allongement du /R/ initial ni la prothèse, ses /R/ initiaux demeurent brefs et apicaux. Enfin, en ce qui concerne les codas, chez les locuteurs disposant d'une opposition de longueur transcrite en opposition [ʁ] ~ [r, ɾ], ce /R/ en coda ne saurait être phonologiquement long et donc phonétiquement postérieur, tandis que les autres locuteurs n'ont que des [r, ɾ] antérieurs.

En francoprovençal, les deux contextes où l'on observe le moins de [r] apicaux sont les frontières de mots (46 % à l'initiale et 32 % en finale). Les autres positions forment une classe relativement homogène et plus proche des résultats des autres dialectes, à 73 % de [r] apicaux.

Dans le domaine d'oïl, la coda interne présente le plus faible pourcentage de réalisations apicales (65 % en bourguignon, 67 % en mainiot), tandis que toutes les autres positions aboutissent à des taux supérieurs à 86 % en bourguignon, croissant régulièrement en mainiot (de 71 à 93 %). Les différences de détails entre ces deux dialectes sont à prendre avec précaution, compte tenu du faible nombre d'observations dans certains contextes.

En bourguignon et dans les différentes variétés d'occitan, la rhotique finale demeure apicale lorsqu'elle est prononcée, mais elle est régulièrement amuïe au niveau lexical (pour les infinitifs, par exemple). Cela montre que, si [ʁ] peut être analysé comme un affaiblissement modéré vis-à-vis de l'amuïssement, il ne saurait être analysé comme une étape nécessaire de ce dernier. De surcroît, la tendance générale à l'affaiblissement en fin de mot demeure sujette à une forte variation selon les dialectes.

# 4 Résultats en fonction du contexte sociolinguistique

Nos locuteurs auront, en 2018, entre 38 et 88 ans. Il s'agit ici (et dans la fig. 2) d'étudier la relation entre leur âge et la variation de leur /R/. L'année de naissance, en effet, conditionne notamment le système éducatif et les représentations de la diglossie au contact desquels les enfants ont grandi.

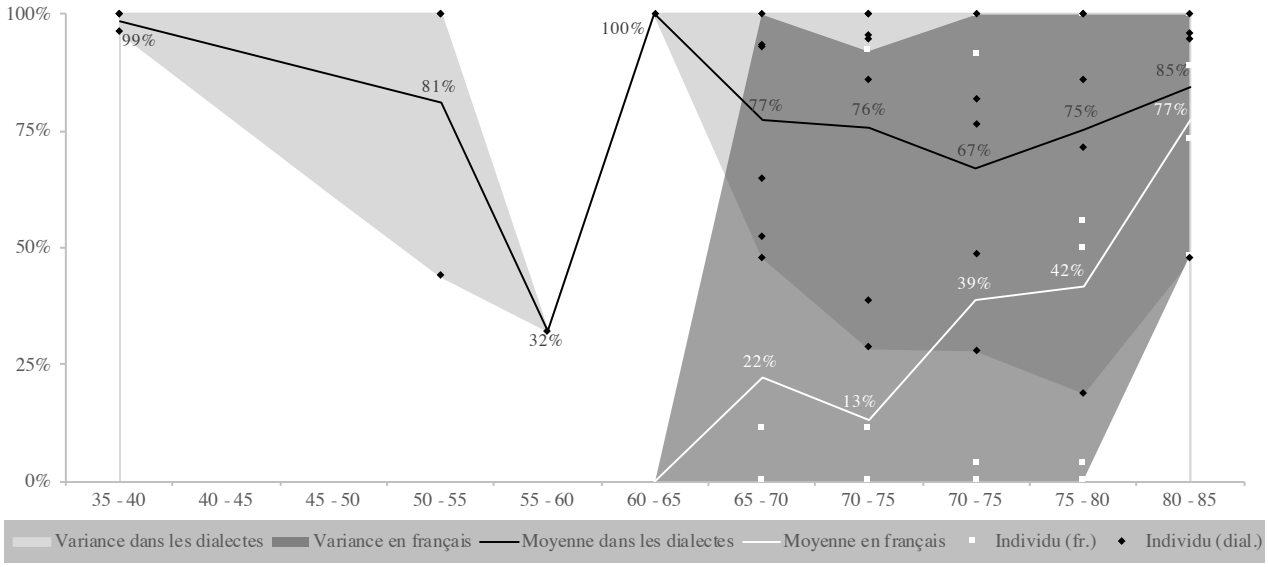



FIGURE 2 – Taux de [r] apicaux en français régionaux et en dialectes. Les surfaces autour des courbes représentent la dispersion des données, du locuteur produisant le moins de [r] apicaux à celui qui en produit le plus, pour chaque tranche d'âge.

## 4.1 En français

Pour les variétés de français régional, le dessin de la courbe blanche (fig. 2) est assez clair : plus un locuteur est âgé, plus il est susceptible de produire des rhotiques antérieures. Cependant, la variance (dont l'étendue est représentée par les zones grises) demeure élevée et cette moyenne n'est qu'une tendance générale.

On notera simplement qu'en dessous de 65 ans, aucun locuteur ne « roule » les *r* en français, le milieu du XX$^{e}$ siècle jouant un rôle de pivot dans la convergence partielle des systèmes phonétiques dialectaux et français (*cf.* 4.3). Au contraire, on notera qu'aucun de nos locuteurs de plus de 85 ans n'est capable de produire 100 % de rhotiques postérieures, les années 1930 marquant un autre pivot. Des années 1930 aux années 1950 s'étire donc un *continuum*, résultat de la régression des patois, de leur connotation négative et de l'affermissement de la norme phonétique enseignée par l'école.

## 4.2 Dans les dialectes et langues minoritaires

Dans les dialectes, la situation est assez différente (fig. 2, courbe noire). Cette différence est notamment due aux locuteurs occitans, qui représentent tous nos locuteurs de moins de 55 ans. Ceux-ci, néo-locuteurs ou ayant appris la langue dans le cercle familial mais avec un certain regard grammatical, ne produisent que peu ou pas de /R/ dorsaux. La valorisation de la langue minoritaire, active chez ces locuteurs, et le cadre culturel reconnu dans lequel elle s'inscrit leur permettent de différencier assez bien les deux systèmes phonétiques et de ne pas réaliser de « *r* français » en occitan. Ce constat vaut pour tous les locuteurs occitans de moins de 55 ans, sauf un. Celui-ci, pourtant artiste occitanophone revendiqué et lié au milieu militant, ne produit que 44 % de [r] apicaux en gascon. De son propre aveu, il « n'arrive pas bien à rouler les *r* » et semble le regretter.

Le taux plus faible de [r] antérieurs dans la tranche de 55 à 60 ans provient d'un seul locuteur,

francoprovençal. Si l'on omet les Occitans, ce taux de 32 % n'est pas si faible et correspond à une chronologie relativement parallèle à celle du français régional. Le profil de ce locuteur, impliqué dans la défense, la promotion et l'enseignement du savoyard, correspond à une volonté partiellement réalisée d'articuler des /R/ non-standards.

Le mouvement qu'amorce ensuite la courbe, de 60 à 87 ans, est encore une fois à considérer avec précaution. Un seul locuteur, en effet, est représenté dans la tranche d'âge de 60 à 65 ans. Chercheur spécialiste de l'occitan, celui-ci dispose d'une aptitude à séparer les deux systèmes phonétiques selon une alternance de type « tout ou rien » (100 % de [r] en occitan et 100 % de [ʁ] en français). Abstraction faite de son cas, la descente de la courbe vers 75–80 ans ne représente que la moyenne d'une situation assez hétéroclite et n'est pas signifiante, étant accompagnée d'un fort écart type (28 % en moyenne, de 65 à 90 ans).

### 4.3 Éléments de diachronie de la diffusion des /R/ postérieurs

La co-présence de deux langues en situation de diglossie peut provoquer des effets de convergence entre les systèmes linguistiques, influant notamment sur la réalisation rhotique (Spreafico & Vietti, 2013). Concernant les variétés de français régional, la fig. 2 permet de visualiser les trois phases suivantes : (1) les locuteurs de 80 ans et plus disposent tous, dans notre corpus, de [r] en français ; ils se situent dans un stade de convergence où la rhotique française est sous l'influence de la rhotique dialectale. (2) Les locuteurs de 65 à 80 ans se situent dans une phase de transition ; leur proportion de [r] en français diminue régulièrement. (3) Les locuteurs de moins de 65 ans ne produisent plus de [r] en français.

Il y a donc eu une transition, dans les conditions d'apprentissage de la langue, entre les locuteurs nés au début des années 1930, susceptibles de produire une certaine proportion de [r] apicaux en français, et les locuteurs nés à partir des années 1950, qui ne produisent plus de [r] apicaux en français. Pour les dialectes, la même conclusion est néanmoins impossible à tirer de nos données, qui ne comptent pas de jeunes locuteurs en dehors du domaine occitan. Cette absence, dans le domaine d'oïl, n'est que le reflet du déclin des langues minoritaires.

Dans tous les dialectes étudiés ici, on trouve des locuteurs produisant uniquement des [r] antérieurs et des locuteurs en produisant très peu. On trouve de plus des locuteurs âgés ne produisant que des [ʁ] d'arrière, hors de notre corpus. Mais celui-ci, par construction, ne les a pas retenus.

## 5 Conclusion et perspectives

Cette étude montre que, dans la diglossie gallo-romane, les dialectes et le français ne constituent pas des systèmes imperméables. L'articulation du /R/, chez la plupart des locuteurs, n'est pas toujours clairement implémentée comme apicale en dialecte et dorsale en français. Les différents systèmes phonétiques présents chez un locuteur donné tendent à communiquer et à se contaminer l'un l'autre. Cette contamination s'étend dans (au moins) deux dimensions, en fonction du contexte phonologique et du profil sociolinguistique du locuteur.

**Variation selon le contexte phonétique.** Au niveau pan-dialectal, il convient non d'opposer simplement les positions d'attaque et de coda, mais d'opposer surtout les positions finales de mots aux autres positions. Dans plusieurs variétés dialectales comme en français régional, la position de finale

montre un comportement spécifique : elle favorise généralement un allophone postérieur. Dans la mesure où cette position est marquée, pour d'autres lexèmes et dans (presque) toutes les variétés gallo-romanes, par l'effacement du /R/, nous analysons ce [ʁ] postérieur comme une variante faible, à mi-chemin entre un /r/ antérieur et une absence de /R/. Cette tendance est nette en français et en dialectes, sur les territoires d'oïl et de francoprovençal, mais ne s'applique pas à l'aire occitane.

Dans certaines variétés de français, de fait, le contexte phonologique n'a que peu ou pas d'influence sur la réalisation de la rhotique ; c'est le cas des variétés languedociennes et gasconnes et, à l'exception de la position finale, des variétés francoprovençales et bourguignonnes. En français du Maine, au contraire, c'est la position initiale qui présente le taux le plus bas de [ʁ] dorsaux. En ce qui concerne les dialectes et langues minoritaires, deux d'entre eux montrent une importance spécifique des positions de frontière de mot. En francoprovençal, ces deux positions donnent lieu à moins de [r] apicaux que les autres contextes ; en gascon, le phénomène est inversé et toutes les rhotiques bordées par une frontière de mot sont prononcées [r].

**Variation selon le profil sociolinguistique.** Chez les locuteurs dont le dialecte a été la langue dominante, le [r] apical tend à contaminer largement le français régional. Au contraire, chez les locuteurs qui ont appris le dialecte dans une situation déjà minoritaire et qui ont subi l'idéologie normalisatrice du français, c'est le [ʁ] qui a contaminé leur dialecte. Ce paradigme, néanmoins, n'est pas universel : le degré de surveillance et le cadre de l'apprentissage influent fortement sur la capacité des locuteurs à différencier les deux grammaires.

**Limites de l'étude et perspectives.** D'autres facteurs interviennent et pourraient nuancer nos analyses. Un travail portant sur la parole spontanée produirait un précieux contrepoint à cette étude. En conversation bilingue, par exemple, le [ʁ] français peut fortement pénétrer le dialecte, comme le note l'ALJA (Martin & Tuaillon, 1978). Ce phénomène, pendant synchronique de la convergence mentionnée plus haut, se retrouve dans notre corpus de parole spontanée dans des alternance codiques (*code switching*) qui demanderaient à être documentées. Par ailleurs, une étude de l'allophonie plus détaillée, dépassant l'opposition apicale/dorsale et appuyée sur des données articulatoires, offrirait une image plus précise des différentes dynamiques du /R/ gallo-roman. Enfin, une utilisation de profils sociologiques plus complets et sur un échantillon plus large, intégrant notamment le sexe, la profession et une interrogation des représentations linguistique, permettrait d'affiner l'analyse sociolinguistique. Malgré ses limites, cette étude présente des données inédites sur la distribution de [r] apicaux et [ʁ] dorsaux dans quelques dialectes gallo-romans et variétés de français régional, dessinant le paysage suivant : la diffusion du français et le bilinguisme qui s'est ensuivi ont entraîné une contamination du [r] dialectal par le [ʁ] du français standard.

De par sa fréquence et le nombre d'allophones qu'il présente, ce phonème demeure un indice fort de perception d'un *accent* régional, social ou étranger (Boula de Mareüil, 2010). Au-delà du développement en diachronie du /R/ et de sa variation diatopique (géographique), une étude de la variation diastratique et diaphasique (socioculturelle et stylistique) reste nécessaire.

# Remerciements

Ce travail a en partie été financé par la Délégation Générale à la Langue Française et aux Langues de France (DGLFLF), dans le cadre du programme « Langues et numérique » 2016. Nous remercions Michela Russo pour sa contribution, ainsi que tous les locuteurs qui ont rendu possible ce travail.

# Références


BOISGONTIER J., MICHEL L., RAVIER X. & PETIT J.-M. (1981). *Atlas linguistique et ethnographique du Languedoc oriental (ALLOr)*, volume I. Paris : éd. du CNRS.

BOULA DE MAREÜIL P. (2010). *D'où viennent les accents régionaux ?* Paris : Le Pommier.

BOULA DE MAREÜIL P., RILLIARD A. & VERNIER F. (2018). Enregistrements et transcription pour un atlas sonore des langues régionales de france. *Geolinguistique*, **17**, 23–48. Atlas accessible à l'URL `https://atlas.limsi.fr`.

DELL F. (1976). Schwa pécédé d'un groupe obstruante-liquide. *Recherches Linguistiques*, **4**, 75–111.

DEMOLIN D. (1999). Some phonetic and phonological observations concerning /R/ in Belgian French. In H. VAN DE VELDE & R. VAN HOUT, Eds., *'R-atics : Sociolinguistic, phonetic and phonological characteristics of /r/*, p. 63–73. Bruxelles : Institut des Langues Vivantes de et Phonétique.

ENGSTRAND O., FRID J. & LINDBLOM B. (2007). A Perceptual Bridge Between Coronal and Dorsal /r/. In M.-J. SOLE, P. BEDDOR & M. OHALA, Eds., *Experimental Approaches to Phonology*, p. 175–191. Oxford : Oxford University Press.

GARDETTE P. (1990). *Atlas linguistique et ethnographique du Lyonnais (ALLy)*, volume I. Lyon : Institut de linguistique des Facultés catholiques de Lyon.

GENDROT C. (2014). Perception et réalisation du /r/ standard français en finale de mot. In *Actes des journées d'étude de la parole*, Le Mans (http ://www-lium.univ-lemans.fr/jep2014/programme.php).

GILLIÉRON J. & EDMONT E. (1902–1910). *Atlas linguistique de la France (ALF)*. Paris : Honoré Champion.

GUILLAUME G., CHAUVEAU J.-P. & LAGRANGE-BARRETEAU R. (1975). *Atlas linguistique et ethnographique de la Bretagne romane, de l'Anjou et du Maine (ALBRAM)*, volume I. Paris : éd. du CNRS.

LABOV W. (1972). *Sociolinguistic patterns*. Philadelphia : University of Pennsylvania Press.

LODGE A. (1997). *Le Français, Histoire d'un dialecte devenu langue*. Paris : Fayard.

MARTIN J.-B. & TUAILLON G. (1978). *Atlas linguistique du Jura et des Alpes (ALJA)*, volume I. Paris : éd. du CNRS.

MORIN Y. C. (2013). From apical [r] to uvular [ʁ] : what the apico-dorsal *r* in Montreal French reveals about abrupt sound changes. In F. SÁNCHEZ MIRET & D. RECASENS, Eds., *Studies in phonetics, phonology and sound change in Romance*. Munich : LINCOM Europa.

OLIVIÉRI M. & SAUZET P. (2016). Southern Gallo-Romance (Occitan). In A. LEDGEWAY & M. MAIDEN, Eds., *The Oxford Guide to te Romance Languages*, p. 319–349. Oxford : Oxford University Press.

RAVIER X., BOISGONTIER J. & NÈGRE E. (1978). *Atlas linguistique et ethnographique du Languedoc occidental (ALLOc)*, volume I. Paris : éd. du CNRS.

ROUILLÉ N. (2008). *La prononciation de la langue publique aux XVIIème et XVIIIème siècles*. Sampzon : Éditions Delatour France.

SPREAFICO L. & VIETTI A. (2013). On rhotics in a bilingual community : a preliminary UTI research. In L. SPREAFICO & A. VIETTI, Eds., *Rhotics : New Data and Perspectives*, p. 57 – 77. Bozen – Bolzano : Bozen – Bolzano University Press.

STRAKA G. (1979). L'histoire de la consonne *r* en français. In *Les Sons et les mots*, p. 465–499. Paris : Librairie C. Klincksieck.

TAVERDET G. (1975). *Atlas linguistique et ethnographique de Bourgogne (ALB)*, volume I. Paris : éd. du CNRS.

# Quand les voyelles longues et brèves ne tiennent pas en place : la qualité vocalique en allemand L2

Jane Wottawa[1, 2] Martine Adda-Decker[2]
(1) LIMSI, CNRS, Université Paris-Saclay,
Bât 508, rue John von Neumann, Campus Universitaire, 91405 Orsay, France
(2) LPP, UMR 7018 CNRS - U. Paris 3 / Sorbonne Nouvelle,
19 rue des Bernardins, 75005 Paris, France
jane.wottawa@univ-paris3.fr, martine.adda-decker@univ-paris3.fr

## RÉSUMÉ

En allemand, la distinction entre voyelles longues et brèves n'est pas seulement liée à une différence de durée mais aussi à une différence de qualité dans la plupart des cas. Nous examinons dans quelle mesure les apprenants francophones de l'allemand réalisent une différence de qualité entre les voyelles longues et brèves en comparaison avec des germanophones natifs. Les résultats montrent que de manière générale, les apprenants francophones de l'allemand produisent des voyelles avec un F2 plus élevé que les germanophones natifs. Dans l'espace vocalique, les voyelles peuvent être qualifiées comme étant plus antérieures. Seulement pour la paire [iː, ɪ] les apprenants de l'allemand et les germanophones natifs se distinguent sur les valeurs des trois premiers formants. La différence entre [iː] et [ɪ] est bien marquée pour les deux premiers formants chez les germanophones natifs. Les productions des apprenants de l'allemand montrent plutôt des différences sur le F3 pour cette paire vocalique.

## ABSTRACT

**When short and long vowels do not keep in place : vowel quality in German L2.**

In German, the distinction between long and short vowels is not only related to a difference in duration but also to a difference in quality most of the time. We investigated to what extend French learners of German realize a quality difference between long and short vowels compared to German native speakers. In general, French learners of German produce vowels with a higher F2 than German native speakers. In the vowel space, vowels can be qualified as more fronted. Only for [iː, ɪ] learners of German and German native speakers differ in the values of the first three formants. The difference between [iː] and [ɪ] is highly marked for the first two formants in German native speakers. The learners' productions rather show differences in F3 for this vowel pair.

MOTS-CLÉS : production des voyelles L2; allemand langue étrangère; français langue maternelle; qualité vocalique.

KEYWORDS: L2 vowel production, German L2, French L1, vowel quality.



10.21437/JEP.2018-8

# 1 Introduction

La prononciation en langue étrangère peut différer de la prononciation d'un locuteur natif de la même langue. Nous nous intéressons ici à la prononciation de quelques voyelles longues et brèves allemandes par des germanophones natifs et des apprenants francophones de l'allemand.

A notre connaissance, une étude comparative de la qualité vocalique chez les apprenants francophones de l'allemand et des germanophones natifs n'a pas encore été menée. Cependant, les résultats d'un test de perception réalisé auprès de germanophones natifs amenés à identifier des mots produits par des apprenants francophones de l'allemand ont été publiés (Zimmerer & Trouvain, 2015). Ce test de perception révèle que les germanophones natifs sont en effet capables de correctement identifier les voyelles produits par les apprenants et cela d'autant plus que les apprenants sont d'un niveau avancé en allemand. En revanche, les voyelles brèves ont été moins souvent identifiées correctement que les voyelles longues. Cette tendance était observée à la fois pour les productions des apprenants débutants et des apprenants avancés.

Dans une étude préalable sur le même corpus qui est étudié dans l'étude présente, nous avons observé que les apprenants francophones de l'allemand n'ont pas de difficultés à produire la différence de durée des voyelles longues et brèves (Wottawa *et al.*, 2018). Nous voulons maintenant analyser la réalisation de la qualité vocalique entre voyelles longues et brèves chez les apprenants francophones de l'allemand et les comparer à des productions chez des germanophones natifs. Pour ce faire, nous avons d'abord comparé les distances euclidiennes des paires vocaliques entre germanophones natifs et des apprenants de l'allemand. Une comparaison similaire a été effectuée entre des locuteurs natifs de l'anglais américain et des apprenants mandarins de l'anglais (Chen, 2006). Ensuite, nous avons comparé les voyelles par paire longue-courte afin de déterminer leurs positions différentes dans l'espace vocalique.

# 2 Méthodes

Dans cette section, nous présentons le corpus (FLACGS) et les analyses acoustiques, incluant le calcul de distances euclidiennes entre voyelles ainsi qu'une comparaison des trois premiers formants.

## 2.1 Corpus de parole et matériel acoustique

Les analyses acoustiques ont été menées sur le *French Learners' Audio Corpus of German Speech* (FLACGS) Corpus (Wottawa & Adda-Decker, 2016). Le corpus contient des enregistrements de 40 locuteurs en allemand dont 20 locuteurs germanophones natifs et 20 apprenants de l'allemand ayant le français comme langue maternelle. Le niveau de compétences des apprenants allait de A2 à C2 avec la distribution suivante : A1/A2 : 3 participants, B1/B2 : 9 participants, C1/C2 : 8 participants.

Dans le cadre de cette étude, deux tâches de production ont été analysés : la répétition de mots dans des phrases cadre et la lecture de deux textes allemands (*Nordwind und Sonne, Die Buttergeschichte*). Dans l'ensemble, le volume de production représente un peu plus que quatre heures de parole. Le corpus a été transcrit manuellement et aligné automatiquement par `web-Maus` (Kisler *et al.*, 2017). Pour les voyelles, l'alignement automatique a été corrigé manuellement si nécessaire en se basant sur le deuxième formant et ses mouvements. Si la voyelle était suivie d'une consonne voisée ou d'une

autre voyelle, le choix des frontières de segment était guidé par les mouvements du F2 et de l'intensité relative. La forme du signal acoustique fournissait également quelques indications sur l'emplacement des frontières. Seules les voyelles apparaissant dans des syllabes portant l'accent lexical entraient dans l'analyse des voyelles longues et brèves.

En utilisant Praat (Boersma & Weenink, 2016), les valeurs formantiques en Hertz des trois premiers formants ont été extraites toutes les 5 ms à partir des enregistrements audio. Ensuite, elles étaient moyennées pour l'ensemble de la voyelle ainsi que pour le début, le centre et la fin de la voyelle. Pour les analyses décrites par la suite, nous avons utilisé uniquement les valeurs du centre des voyelles transformées en Bark.

## 2.2 Analyses acoustiques

### 2.2.1 Distances euclidiennes

Nous avons calculé les distances euclidiennes entre les voyelles suivantes : [iː, ɪ], [yː, ʏ], [aː, a] et [oː, ɔ] pour chaque participant. Dans un premier temps, ces distances incluaient uniquement les distances correspondant à F1 et F2 (Traunmüller, 1981). Dans un deuxième temps, nous avons ajouté également la distance euclidienne pour F3.

D'abord, nous avons comparé les distances euclidiennes calculées avec les deux premiers formants à celle calculées avec les trois premiers formants. Ensuite, nous avons établi les relations entre les distances euclidiennes et le groupe de locuteurs (germanophones natifs, apprenants de l'allemand) et les tâches de production (répétition, lecture). Pour ce faire, nous avons eu recours modèle mixte à pente aléatoire (random slope model) (Winter, 2013) en utilisant la *library* `lme4` (Bates *et al.*, 2015) intégré en R (R Development Core Team, 2008).

### 2.2.2 Comparaison des valeurs formantiques

La distance euclidienne entre les voyelles longues et brèves permet de quantifier les différences de production de la qualité vocalique entre germanophones natifs et apprenants de l'allemand. Cependant, nous avons également effectué des régressions logistiques multiples des trois premiers formants pour chacune des paires vocaliques cf. [iː, ɪ], [yː, ʏ], [aː, a], [oː, ɔ] afin de mieux déterminer la position des voyelles dans l'espace vocalique. De cette manière, la localisation des deux membres d'une paire vocalique peut être comparée entre les productions des natifs et celle des apprenants.

# 3 Résultats

Dans cette section, nous présentons les résultats de deux analyses. Dans un premier temps, les distances euclidiennes sont présentées et analysées en utilisant un modèle mixte à pente aléatoire. Ensuite, nous comparons les valeurs formantiques des voyelles entre les production des natifs et les apprenants par le moyen de la régression logistique multiple.

## 3.1 Distances euclidiennes

Tout d'abord, nous avons comparé les distances euclidiennes calculées uniquement avec F1 et F2 à celles calculées avec F1, F2 et F3 en utilisant un test de Student apparié pour chacun des groupes, c'est-à-dire les germanophones natifs et les apprenants de l'allemand et pour chacune des tâches, c'est-à-dire la répétition et la lecture. Les résultats de ces comparaisons ont montré qu'en général, l'ajout de F3 dans le calcul augmente significativement la distance entre les voyelles longues et brèves. Seul pour le contraste [oː, ɔ] chez les germanophones natifs en répétition et les apprenants de l'allemand en lecture, le calcul comportant les valeurs de F3 ne change pas significativement la distance euclidienne. En d'autres termes, ces résultats montrent que dans nos données le F3 joue un rôle important dans la production des voyelles longues et brèves à la fois chez les germanophones natifs et les apprenants de l'allemand. C'est pourquoi les analyses suivantes ont été effectuées avec les distances euclidiennes calculées à partir des trois premiers formants.

Nous avons étudié les relations entre les distances euclidiennes, les groupes de locuteurs (germanophones natifs, apprenants de l'allemand) et les tâches de production (répétition, lecture) à l'aide d'un modèle mixte à pente aléatoire (random slope model). L'importance statistique pour chaque facteur fixe a été calculée par un test du rapport des vraisemblances selon une approche allant d'un modèle simple à un modèle complexe. Les facteurs fixes étaient testés dans l'ordre : Paire vocalique ([iː, ɪ], [yː, ʏ], [aː, a], [oː, ɔ]), Genre du locuteur, Groupe (germanophones natifs, apprenants de l'allemand) et Tâche (répétition, lecture). Nous avons obtenu des *intercepts* pour le facteur aléatoire Participant ainsi que les interactions Paire vocalique par Participant et Tâche par Participant.

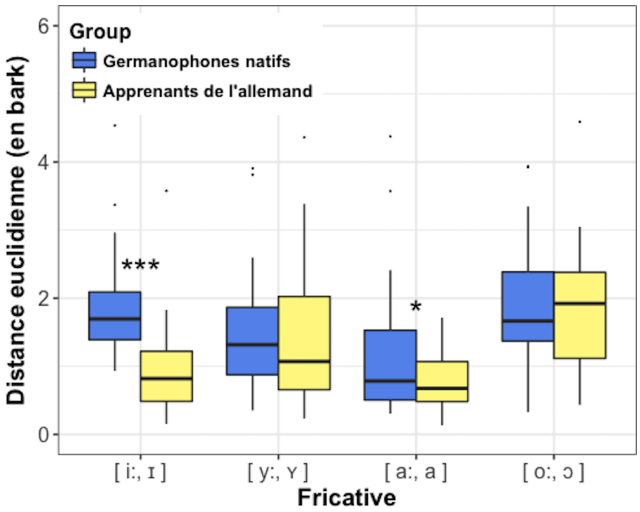



Figure 1: Interaction pour les distances euclidiennes entre Paire vocalique et Groupe ; une différence significative entre groupes est observée pour [iː, ɪ] et [aː, a] ($^{*}$ $p \leq 0.05$ ; $^{**}$ $p \leq 0.01$ ; $^{***}$ $p \leq 0.001$)

Inclure le facteur Genre n'a pas augmenté la précision du modèle le plus simple qui avait comme seul facteur fixe Paire vocalique ($\chi^2(1) = 0.12, p = .73$). En revanche, le modèle a gagné en précision en incluant le facteur Groupe ($\chi^2(1) = 7.94, p < .01$). Chez les germanophones natifs, les distances euclidiennes sont d'environ 0.5 Bark (±0.2 Bark) plus grandes par rapport à l'*intercept* (0.7 Bark (±0.1 Bark)). Ensuite, nous avons testé s'il y a une interaction entre Paire vocalique et Groupe ($\chi^2(3) = 9.83, p < .05$) illustrée par la Figure 1. Par rapport à l'*intercept* (0.8 Bark (±0.2 Bark)), les germanophones natifs produisent une distance euclidienne qui est 0.6 Bark (±0.3 Bark) plus grande pour la paire vocalique [iː, ɪ], et 0.5 Bark (±0.4 Bark) plus petite pour la paire vocalique [yː,

ʏ] et de 0.2 Bark (±0.4 Bark) plus petite pour la paire vocalique [oː, ɔ]. Des comparaisons *post-hoc* ont révélées que la différence des distances euclidiennes entre Groupe n'est significative que pour les paires vocaliques [iː, ɪ] (t(77) = 4.3, p < .001) et [aː, a] (t(54) = 2.5, p < .05). Dans les deux cas, les germanophones natifs ([iː, ɪ] : *M*=2.0 Bark (±0.9 Bark), [aː, a] : *M*=1.2 Bark (±0.88 Bark)) montrent une distance euclidienne plus importante que les apprenants de l'allemand ([iː, ɪ] : *M*=1.8 Bark (±1.2 Bark), [aː, a] : *M*=0.8 Bark (±0.4 Bark)). Finalement, inclure le facteur Tâche ne sert pas à la précision du modèle ($\chi^2(1) = 1.31, p = .25$) c'est pourquoi nous l'avons écarté de l'analyse.

## 3.2 Comparaison des valeurs formantiques

La distance euclidienne entre les voyelles longues et brèves permet de quantifier les différences de production de la qualité vocalique. Cependant, elles ne nous informent pas sur la position des voyelles dans l'espace vocalique. C'est pourquoi nous avons mené une analyse des trois premiers formants par rapport au Groupe (germanophones natifs, apprenants de l'allemand) pour chacune des paires vocaliques, cf. [iː, ɪ], [yː, ʏ], [aː, a] et [oː, ɔ]. Le Groupe était fixé comme la variable dépendante de la régression logistique multiple. Les variables Voyelle, Genre du locuteur, F1, F2, F3 et Tâche (répétition, lecture) étaient les variables explicatives testées dans cet ordre en allant du modèle le plus simple au modèle le plus complexe. Le modèle nul ne comportait que la variable explicative Voyelle.

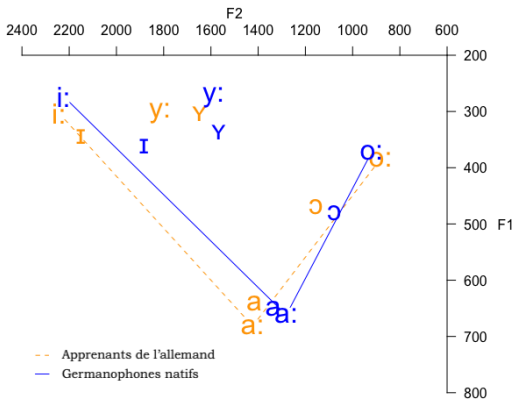



Figure 2: Interaction entre Paire vocalique et F2 par Groupe, espace vocalique F1-F2 en Hz

Les variables Genre de locuteur et Tâche n'ont pas été statistiquement significatives dans les calculs et cela pour aucune paire vocalique. Cependant, les analyses nous informent que pour les paires vocaliques [yː, ʏ], [aː, a] et [oː, ɔ], la différence entre germanophones natifs et apprenants de l'allemand n'est significative que pour les valeurs du F2 et cela d'autant plus si le F2 entre en interaction avec la variable Voyelle ([yː, ʏ] : ($\chi^2(1) = 7.59, p < .01$) ; [aː, a] : ($\chi^2(1) = 17.12, p < .001$) ; [oː, ɔ] : ($\chi^2(1) = 5.27, p < .05$)). Figure 2 illustre ces interactions. Pour la paire [yː, ʏ], les germanophones natifs ([ʏ] : *M*=11.3 Bark (±1.1 Bark) ; [yː] : *M*=11.0 Bark (±3.5 Bark)) ont 1.4 fois plus de chance de produire la voyelle [yː] si le F2 baisse d'un bark que les apprenants de l'allemand ([ʏ] : *M*=11.4 Bark (±2.8 Bark) ; [yː] : *M*=12.3 Bark (±1.1 Bark)). Pour la paire [aː, a], les germanophones natifs ([a] : *M*=10.0 Bark (±3.1 Bark) ; [aː] : *M*=9.8 Bark (±2.3 Bark)) ont 1.3 fois plus de chance de produire la voyelle [aː] si le F2 baisse d'un bark que les apprenants de l'allemand ([a] : *M*=10.6 Bark (±1.7 Bark) ; [aː] : *M*=10.7 Bark (±1.2 Bark)). Pour la paire [oː, o], les germanophones natifs ([o] : *M*=8.8 Bark (±2.0 Bark) ; [oː] : *M*=7.8 Bark (±2.1 Bark)) ont 1.2 fois plus de chance de produire la voyelle [oː] si le F2 augmente d'un bark que les apprenants de l'allemand ([o] : *M*=9.3 Bark (±1.4

Bark) ; [oː] : *M*=7.7 Bark (±1.4 Bark)).

Outre F2, F3 permet de distinguer la paire vocalique [aːˑ a] entre germanophones natifs et apprenants ($\chi^2(1) = 9.45, p < .05$)). Les germanophones natifs ([a] : *M*=13.6 Bark (±4.1 Bark) ; [aː] : *M*=14.4 Bark (±3.0 Bark)) ont 1,3 fois plus de chance de produire [aː] si F3 augmente d'un bark que les apprenants de l'allemand ([a] : *M*=14.6 Bark (±2.0 Bark) ; [aː] : *M*=14.9 Bark (±1.0 Bark)). Les Groupes (germanophones natifs, apprenants de l'allemand) produisent [iː] et [ɪ] avec des valeurs différentes pour chacun des trois premiers formants (F1 : $\chi^2(1) = 28.42, p < .001$) ; F2 : $\chi^2(3) = 75.47, p < .001$) ; F3 : $\chi^2(7) = 54.04, p < .001$). Les moyennes pour [iː] et [ɪ] sont résumés dans Table 1. Le tableau et les résultats des analyses indiquent que par rapport aux apprenants de l'allemand, les germanophones ont 2.1 fois plus de chances de produire [iː] si le F1 baisse d'un bark, 1.3 fois plus de chances de produire [iː] si le F2 augmente d'un bark et 4.7 fois plus de chances de produire [iː] si le F3 baisse d'un bark.

| Groupe | Voyelle | F1 (bark) | | F2 (bark) | | F3 (bark) | |
|---|---|---|---|---|---|---|---|
| | | *M* | ± | *M* | ± | *M* | ± |
| germanophones natifs | [ɪ] | 3.6 | 0.9 | 12.4 | 2.2 | 15.0 | 2.3 |
| | [iː] | 2.7 | 0.9 | 13.3 | 3.2 | 15.2 | 3.6 |
| apprenants de l'allemand | [ɪ] | 3.4 | 0.8 | 13.3 | 1.8 | 15.5 | 1.7 |
| | [iː] | 3.0 | 0.7 | 13.7 | 1.4 | 16.0 | 1.3 |

Table 1: Valeurs moyennes pour F1, F2 et F3 par Groupe en bark

# 4 Discussion

Tout d'abord, nous avons calculé les distances euclidienne pour les paires vocaliques [iː, ɪ], [yː, ʏ], [aː, a] et [oː, ɔ]. Dans un premier temps, nous avons comparé les distances euclidiennes incluant F1 et F2 à des distances euclidiennes comportant F1, F2 et F3. Nos résultats ont montré que les distances sont plus grandes lorsque le F3 est inclus dans le calcul. Ce résultat peut être expliqué par le fait qu'en allemand les voyelles longues sont soit plus arrondies soit plus étirées que les voyelles courtes. En d'autres termes quant au F3, les voyelles longues se trouvent plus aux extrémités de l'espace vocalique que les voyelles brèves qui sont plus centrales.

Dans un deuxième temps, nous avons étudié les effets des Groupes (germanophones natifs, apprenants de l'allemand), des Tâches de production (répétition, lecture) sur les distances euclidiennes calculées pour chaque participant pour les paires vocaliques [iː, ɪ], [yː, ʏ], [aː, a] et [oː, ɔ]. En général, les germanophones natifs montrent des distances euclidiennes plus grandes que les francophones. Une analyse par paire vocalique a relevé que les distances euclidiennes des germanophones natifs ne sont significativement plus importantes que pour les paires vocaliques [iː, ɪ] et [aː, a]. Ces résultats sont un premier indice des différences qui existent entre les productions vocaliques des germanophones natifs et des apprenants de l'allemand. Les voyelles [iː] et [ɪ] ainsi que [aː] et [a] semblent moins bien séparées chez les apprenants de l'allemand que chez les germanophones natifs. Le genre des locuteurs et la tâche n'avaient pas d'influence sur les distances euclidiennes. En d'autres termes, les locuteurs féminins et masculins appartenant à un des deux groupes produisent des distances similaires. Les distances ne sont pas affectées par la tâche ce qui suggère qu'aucune des deux tâches est plus facile ou plus complexe pour la production des voyelles en langue étrangère.

Les régressions logistiques ont été menées dans l'objectif de mieux caractériser les différences de la qualité vocalique qui existent entre les productions des locuteurs natifs et des apprenants. Pour ces analyses aussi, le genre des locuteurs et la tâche de production n'ont pas influencé la production des voyelles entre les deux groupes. Que le genre ne joue pas de rôle significatif peut être expliqué par le fait que le nombre d'hommes et de femmes enregistrés était équivalent dans les deux groupes.

Les analyses de régression logistique ont relevé que les productions des deux groupes se distinguent surtout par les valeurs de F2. Les tendances suivantes se profilent dans nos résultats : pour les paires vocaliques [iː, ɪ] et [oː, ɔ], le F2 des deux groupes est comparable pour les voyelles longues. Cependant, pour les paires vocaliques [yː, y] et [aː, a] le F2 des deux groupes est comparable pour les voyelles brèves. Les germanophones natifs produisent la voyelle [ɪ] bien plus centrale que les apprenants dont le F2 de leurs productions de [ɪ] est presque identique à leurs productions de [iː], donc antérieur. Dans l'espace vocalique, les productions du [yː], [aː], [a] et [ɔ] des apprenants occupent également une position plus antérieure que celle des germanophones natifs. De plus, pour les paires centrales [yː, y] et [aː, a], les apprenants francophones ont tendance à produire les voyelles longues plus antérieures que leurs contre-parties brèves. Chez les germanophones natifs, les voyelles centrales tendues sont légèrement plus postérieures que leurs contre-parties relâchées.

Seulement deux paires vocaliques ont montré des différences significatives entre les deux groupes quant aux valeurs du F3 : [iː, ɪ] et [aː, a]. Les valeurs de F3 sont plus élevées pour [iː] que pour [ɪ] chez les apprenants francophones de l'allemand. Cependant, les valeurs de F3 des deux voyelles sont presque identiques chez les germanophones natifs. Il est possible que les apprenants francophones de l'allemand soient conscient de la différence entre [iː] et [ɪ] et essaient de la marquer en produisant un [ɪ] qui est moins étiré que [iː]. Leurs productions de [aː] et [a] montrent des valeurs de F3 quasiment identiques alors que chez les germanophones natifs, le F3 est plus élevé pour [aː] que pour [a]. Les différences du F3 entre les deux groupes pour la paire vocalique [aː, a] pourrait expliquer les différences des distances euclidiennes présentées plus haut.

Seule la paire vocalique [iː, ɪ] permet de distinguer les locuteurs des deux groupes par le F1. En moyenne, chez les germanophones natifs le F1 diffère d'un bark entre [iː] et [ɪ]. Chez les apprenants, la différence n'est que de 0.4 Bark. Dans l'espace vocalique des apprenants de l'allemand, la place restreinte occupée par [iː] et [ɪ] permet une place relativement antérieure pour la voyelle [yː]. Chez les germanophones natifs, au contraire, la distance entre [iː] et [ɪ] est grande engageant surtout les deux premiers formants. En revanche,la distance entre [yː] et [ʏ] est plus faible surtout concernant les valeurs du F2.

# 5 Conclusion

Nos analyses nous ont permis de caractériser les différences de production des voyelles longues et brèves de l'allemand ([iː, ɪ], [yː, ʏ], [aː, a], [oː, ɔ]) entre les locuteurs germanophones natifs et les apprenants de l'allemand enregistrés dans le corpus FLACGS. Cette comparaison systématique des distances euclidiennes et des analyses formantiques en fonction des deux groupes de locuteurs a indiqué que la paire vocalique [iː, ɪ] présente les différences acoustiques les plus importantes entre les deux groupes de locuteurs. Les différences concernent à la fois la distance euclidienne qui est plus petite chez les apprenants de l'allemand que chez les germanophones natifs et les valeurs des trois premiers formants. Cependant, dans l'ensemble des paires vocaliques choisies, la paire [iː, ɪ] est quelque peu atypique. Pour les autres paires vocaliques, les deux groupes de locuteurs se différencient

surtout par les valeurs du F2. Les apprenants francophones ont tendance à produire des voyelles avec des valeurs de F2 plus élevées que les germanophones natifs.

Les résultats de nos analyses acoustiques ne permettent malheureusement pas d'expliquer les résultats du test de perception réalisé auprès de germanophones natifs (Zimmerer & Trouvain, 2015). Ce test indiquait que les voyelles [ʏ] et [ɔ] étaient les moins bien identifiées alors que dans notre corpus, les voyelles natives et non-natives [ʏ] et [ɔ] sont relativement proches et surtout bien distinctes de [yː] et [oː]. Il serait intéressant de mener les mêmes analyses acoustiques présentées en haut sur les stimuli utilisés lors du test de perception afin de comparer ces productions aux productions analysées ici.

# Remerciements

Ce travail a été soutenu par le programme Investissements d'Avenir - Labex EFL (ANR-10-LABX-0083).

# References

BATES D., MÄCHLER M., BOLKER B. & WALKER S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1), 1–48.

BOERSMA P. & WEENINK D. (2016). *Praat: doing phonetics by computer [Computer program]*. Version 6.0.19, retrieved 13 June 2016.

CHEN Y. (2006). Production of tense-lax contrast by Mandarin speakers of English. *Folia phoniatrica et logopaedica*, **58**(4), 240–249.

KISLER T., REICHEL U. & SCHIEL F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, **45**, 326 – 347.

R DEVELOPMENT CORE TEAM (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

TRAUNMÜLLER H. (1981). Perceptual dimension of openness in vowels. *The Journal of the Acoustical Society of America*, **69**(5), 1465–1475.

WINTER B. (2013). A very basic tutorial for performing linear mixed effects analyses. *arXiv preprint arXiv:1308.5499*.

WOTTAWA J. & ADDA-DECKER M. (2016). French Learners Audio Corpus of German Speech (FLACGS). In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, Portorož (Slovenia).

WOTTAWA J., ADDA-DECKER M. & ISEL F. (2018). The impact of production complexity in German L2 by French native speakers: Focus on /h/ and vowel duration contrast. In E. BABATSOULI & D. INGRAM, Eds., *Phonology in Protolanguage and Interlanguage*. Equinox.

ZIMMERER F. & TROUVAIN J. (2015). Perception of French speakers' German vowels. *Interspeech 2015*.

# Étude acoustique de voyelles tenues produites par des patients glossectomisés suite à un cancer endo-bucal

Hasna Zaouali[1], Béatrice Vaxelaire[1], Christian Debry[2], Guy Bronner[3], Rudolph Sock[1&4]

[1]E.A. 1339-Linguistique, Langues et Parole (LiLPa) –ER Parole et Cognition
Institut de Phonétique de Strasbourg (IPS) – Université de Strasbourg
22 rue Descartes – 67084 Strasbourg– Cedex, France.
[2]Service O.R.L. - Hôpitaux Universitaires de Strasbourg
1 av. Molière – 67098 Strasbourg – Cedex, France.
[3]Service O.R.L. - Groupe Hospitalier Saint Vincent-Clinique Sainte Barbe
29 Rue du Faubourg national, 67000 Strasbourg – Cedex, France.
[4]LICOLAB Université Pavla Jozefa Safarika, Faculté des Lettres Košice – Slovaquie.

*hasnazaouali@live.fr*

## RESUME

Le présent travail est une étude acoustique de quelques caractéristiques spectrales de voyelles tenues [i-a-u], produites par des patients ayant subi une glossectomie partielle ou totale, suite à un cancer endo-buccal. Les conséquences de l'opération chirurgicale sont évaluées dans le but d'identifier les différentes perturbations que cette opération peut entraîner, mais également afin de mettre au jour les possibles stratégies de compensations et/ou de réajustements que le patient peut mettre en place seul ou à l'aide d'une rééducation orthophonique. Notre étude est longitudinale, puisque les patients sont enregistrés lors de différentes phases pré et post-opératoires.

## ABSTRACT

**An acoustic study of substained vowels produced by glossectomised patients following endo-oral cancer**

The present study is based on acoustic analyses of some spectral characteristics of sustained vowels [i-a-u], produced by patients having undergone partial or total glossectomy, following endo-oral cancer. The consequences of surgery are evaluated to identify the various perturbations which may be caused by this surgery, as well as to uncover probable *compensatory* or *readjustment strategies* the patient might deploy, alone or with the help of speech therapy. This is a longitudinal investigation since the patients are recorded during various pre and post-surgery phases.

MOTS-CLES : glossectomie, perturbation, réajustement, voyelles tenues, formants F1/F2, l'aire de l'espace vocalique, indice (PHi).

KEYWORDS : glossectomy, perturbation, readjustment, sustained vowels, formants F1/F2, vowel space area, (PHi) index



10.21437/JEP.2018-9

# Introduction

Ce travail s'insère dans le cadre spécifique des problématiques liées aux chirurgies endo-buccales, avec leurs conséquences potentielles au niveau de la production et de la perception de la parole. En effet, les cancers de la cavité buccale, du pharynx et du larynx représentent 12% de tous les cancers, et la France est au premier rang mondial de ces cancers dits des voies aérodigestives supérieures (VADS). Chaque année 650000 nouveaux cas de cancers ORL sont diagnostiqués dans le monde et le plus commun parmi ceux de la cavité orale est celui de la langue (Shin et al., 2012). En effet, la musculature plexiforme de la langue favorise le développement de tumeurs infiltrantes (Meley et Barthelmé, 1987). Le cancer de la langue touche 15000 Français chaque année et entraîne le décès de 5000 patients par an. Dans 95 % des cas, le cancer de la langue est traité par une opération chirurgicale qui consiste à enlever une partie ou la totalité de la langue (glossectomie), du plancher de la bouche (pelvi-glossectomie) et/ou une partie de la mâchoire (pelvi-glosso-mandibulectomie). Ces trois localisations font partie des critères d'inclusion de notre étude. Selon la taille de la tumeur et son extension, ces résections peuvent être partielles ou totales. Par la suite, le chirurgien choisira de former une nouvelle langue par reconstruction par lambeau libre ou local (un muscle prélevé sur une partie du corps : bras, jambe ou d'une région approximante de la zone opérée). Les tumeurs orales et leur traitement peuvent affecter profondément la production de la parole, le goût, la mastication, la sensation lors de la déglutition, et peuvent impacter aussi l'image de soi et la qualité de vie (Crevier-Buchman *et al.*, 2007 ; Savariaux *et al.*, 2001). En dehors du but évident de guérir le patient, ces facteurs doivent être considérés dans la planification du traitement. Le processus de récupération, visant à atteindre les mêmes buts articulatoires et acoustiques qu'avant l'opération, lors desquels les articulateurs sains, mandibule et lèvres, peuvent suppléer la perte de mobilité de la langue, varient d'un patient à un autre. Cela dépend de plusieurs facteurs, parmi lesquels : l'étendue de la lésion (Heller *et al*., 1991 ; Diz Dios *et al*., 1994), l'emplacement de la tumeur (Korpijaakko-Huukha *et al*., 1998), de l'âge des patients et de l'impact de la chirurgie et la reconstruction sur les muscles de la cavité orale (Acher *et al*., 2014 ; Buchaillard *et al*., 2007 et Bressmann *et al*., 2004), les traitements complémentaires en phase post-opératoire, la radiothérapie la curiethérapie, etc. (Shin *et al*., 2012). Les patients doivent être capables d'adapter leur parole à une nouvelle configuration de leur cavité orale.

L'objectif de ce travail est d'évaluer les conséquences d'une glossectomie, partielle ou totale, sur la parole des patients, afin de déceler les différentes perturbations qu'entraine cette opération chirurgicale. Nous souhaitons aussi déceler les possibles stratégies de compensation ou de réajustements que le patient peut mettre en place, seul ou à l'aide d'une rééducation orthophonique, après une exérèse carcinologique plus ou moins importante de la langue. Une approche longitudinale s'avère donc nécessaire. Plus précisément, il s'agit d'analyser les caractéristiques spectrales de la parole de patients souffrant d'un carcinome épidermoïde au niveau de la langue, suivi d'une rééducation orthophonique. L'originalité de notre travail réside : a) dans l'étude de l'indice de dispersion de l'organisation du système vocalique (le PHi), comme moyen de quantifier le réaménagement des espaces vocaliques en fonction des traitements, de la réhabilitation et du temps ; b) dans le fait que nous tâchons de rationaliser nos données dans le cadre du paradigme de la perturbation et des réajustements en production et en perception de la parole (Vaxelaire, 2007).

# 1 Procédure expérimentale

## 1.1 Participants

Cette étude est conduite à partir de données obtenues auprès de 10 patients (3 femmes et 7 hommes) et 3 volontaires sains, appariés en âge et en genre. Tous les patients ont donné leur consentement écrit pour leur participation à l'étude. Tous les participants avaient une vision normale ou corrigée, étaient de langue maternelle française, et n'avaient aucun antécédent de troubles du langage, d'audition, de déficit neurologique ou de pathologie. Différents types de chirurgies ont été réalisés en fonction de la localisation de la tumeur cancéreuse (*cf*. tableau 1).

| Identification patients | Age | Sexe | Profession | TNM | Type d'exérèse | Reconstruction | Radiothérapie | Rééducation ortho |
|---|---|---|---|---|---|---|---|---|
| SIB | 42 | M | Ingénieur | T2N0M0 | Glossect-Partielle G | Non | Non | Non |
| SOM | 56 | M | Ouvrier | T1N0M0 | Glossect-Partielle D | Non | Non | Non |
| ZIM | 51 | F | Manager | T1N0M0 | Glossect-Partielle G | Non | Non | Non |
| GLAD | 52 | M | Conducteur | T2N0M0 | Glossect-Partielle D | Non | Oui | Oui |
| PETR | 64 | F | Retraité | T4N0M0 | Pelvi-Glossect | Non | Oui | Oui |
| JCT | 53 | M | Manager | T4N0M0 | Glossect -Totale | Lambeau libre Antero-lateral | Oui | Oui |
| HACH | 21 | F | Etudiante | T2N0M0 | Hémi-glossectomie | Lambeau libre Antebrachial | Oui | Oui |
| BIRL | 44 | M | Fonctionnaire | T3N0M0 | Pelvi-glosso-mandibulect | Lambeau libre | Oui | Oui |
| ANT | 62 | M | Sans | T2N0M0 | Glossect-Partielle D | Non | Oui | Non |
| ROJ | 52 | M | Fonctionnaire | T1N0M0 | Glossect-Partielle D | Non | Oui | Non |

**Tableau 1** : Répartition des exérèses et informations complémentaires concernant la population de patients (TNM : classification de la taille de la tumeur (T), de la présence d'adénopathies (N) et de la présence de métastases (M), M : homme, F : femme, Pr : profession, glossect : glossectomie, G : gauche, D : droite).

## 1.2 Enregistrements

Ce travail de recherche a été mené en collaboration avec trois établissements hospitaliers en région alsacienne. Les services d'Oto-Rhino-Laryngologie et de Chirurgie Plastique et Maxillo-Faciale du CHU d'Hautepierre, du groupe Hospitalier Saint Vincent (Clinique Sainte Barbe) et des Hôpitaux Civils de Colmar (Hôpital Louis Pasteur). Les enregistrements acoustiques se sont déroulés au sein de ces établissements, dans un environnement silencieux.
Les patients donnaient leur accord pour que leur parole soit analysée dans le cadre de notre recherche en signant un consentement libre et éclairé, avant de procéder aux enregistrements. Les patients ont été enregistrés dans leur chambre au moment de leur admission hospitalière en préopératoire, et en salle de consultation lors des phases post-opératoires (1, 2, 3). Les données ont été acquises comme suit : Préop, soit avant l'opération, entre 1mois et 1.5 mois après l'intervention chirurgicale (Post-op1), à 3 mois et après tous les traitements complémentaires (Post-op2), à 6 mois (Post-op3), etc. Pour des raisons humaines et logistiques, les conditions d'enregistrements n'étaient pas toujours aussi optimales que nous l'aurions souhaité. En effet, le contexte et les consultations étaient parfois éprouvants pour les patients.

## 1.3 Corpus

Le corpus correspond à des voyelles tenues, /i, a, u/ qui permettent d'explorer l'espace vocalique maximal de chaque locuteur. Il s'agit pour le locuteur de prononcer et de tenir environ 5 secondes la voyelle à produire présentée graphiquement. Chaque voyelle est répétée 10 fois, tout en tenant compte des éventuelles difficultés et du degré de fatigabilité du patient, particulièrement dans les phases d'enregistrement post-opératoire. Les voyelles sont présentées sur des cartons au patient sous

leurs formes graphiques : -i- -a- -ou- pour que le patient soit le plus à l'aise possible avec la lecture du corpus.

## 1.4 Mesures

Les 3 voyelles extrême tenues /i, a, u/ permettent d'explorer les capacités maximales des gestes linguaux de nos sujets, afin de comparer les espaces vocaliques des sujets sains et des sujets pathologiques. Les mesures ont été réalisées à l'aide du logiciel PRAAT©. Les mesures effectuées pour les trois voyelles extrêmes sont les suivantes : 1) les deux premiers formants F1 et F2 ; 2) l'espace vocalique maximal ($kHz^2$) à partir de la formule de Héron ; 3) le (PHi) : l'indice de mesure de dispersion de l'organisation du système vocalique. Il est conçu par analyse analogique de variances en comparant la variabilité inter-catégorie vocalique à la variabilité intra-catégorie vocalique. Nous obtenons les valeurs du (PHi) après application des équations indiquées dans (Huet & Harmegnies, 2000).

## 3. Hypothèses

En phase post-opératoire et suite à la chirurgie, la parole des patients pourrait être altérée et modifiée : (1) Nous avons supposé que la taille de l'exérèse, le site de la tumeur et le type de reconstruction auraient un impact direct sur la parole. Cela se manifesterait par des modifications au niveau des valeurs formantiques (F1/F2) ; (2) Les perturbations formantiques pourraient affecter la taille et la forme de l'espace vocalique ; (3) Une désorganisation devrait apparaître au niveau du système vocalique qui se présenterait alors sous forme de l'élévation et de l'abaissement de l'indice de mesure de dispersion (PHi) ; (4) Le temps et la rééducation devraient permettre une amélioration dans la production de la parole chez les sujets atteints, qui devrait progressivement apparaître dans les phases post-opératoires tardives (Post-op 2, 3) et donc, on devrait constater une normalisation des valeurs des paramètres mesurés. Ces derniers seraient alors comparables aux valeurs mesurées en phase préopératoire.

## 4. Résultats

## 4.1 Valeurs formantiques

Suite à l'hétérogénéité des exérèses subies par chaque patient, en ce qui concerne les valeurs formantiques, nous exposons les résultats de quatre cas représentatifs, dont deux avec reconstructions et deux sans reconstruction : une glossectomie partielle, une pelvi-glossectomie, une pelvi-glosso-mandibulectomie et une glossectomie totale.
De manière globale, les valeurs formantiques de F1 et F2 varient en fonction de la phase d'enregistrement pratiquement pour tous les patients et proportionnellement selon chaque opération qu'a subi chacun d'entre eux. Pour les quatre patients présentés, nous avons observé des modifications des valeurs de F1 et F2 pour les trois voyelles extrêmes [i, a, u], lors des phases d'enregistrement postop1 et 2, avant de retrouver les valeurs attendues en postop 3. Ces variations ont des influences directes sur la taille de l'espace vocalique de chaque patient (*Cf.* figure 1 & 2).
Notons que la taille des espaces vocaliques du patient (Zim) et (Petr), et les configurations de ces espaces vocaliques restent pratiquement les mêmes avec de légères modifications de la synergie linguale en postop1 (en vert) et en postop2 (en violet) par rapport aux autres phases d'enregistrement préop (en rouge) et postop3 (en bleu clair) (*Cf.* figure 1).

(Zim) (Petr)

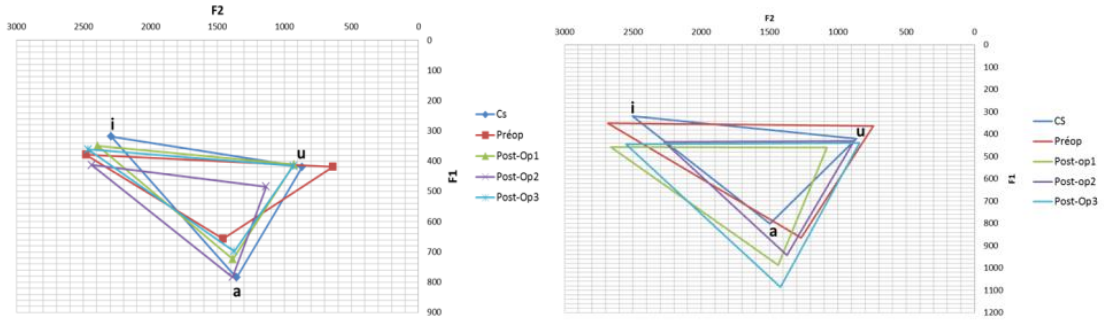



**Figure 1** : Valeurs formantiques de F1 et F2 des voyelles tenues [i,a,u] après une glossectomie partielle (Zim) à gauche et une pelvi-glossectomie (Petr) à droite

Cependant, pour le patient (Birl) ayant subi une pelvi-glosso-mandibulectomie suivie d'une reconstruction avec lambeau libre du péroné droit, nous avons observé que par rapport au préop (le triangle en rouge sur la figure), en postop 1 ce patient a plus de difficulté à réaliser la voyelle [a]. Cela peut être expliqué par le fait que la reconstruction mandibulaire empêche le processus d'ouverture totale de la cavité buccale pour la réalisation correcte et complète de la voyelle [a].
En ce qui concerne le patient (Jct), ayant subi une glossectomie totale suivie d'une reconstruction par un lambeau antéro-latéral de la cuisse droite, nous avons pu observer que par rapport au préop (triangle en rouge), les trois voyelles extrêmes du triangle vocalique sont moins éloignées, et que le triangle est centralisé en phase postop1 (en vert). Ce constat peut être considéré comme une conséquence de la reconstruction, puisque le lambeau est fixé, ce qui réduit sa mobilité. De ce fait, le patient tente de réaliser les voyelles demandées avec un ratage de cible articulatoire, et tout en essayant d'adapter la dynamique de ses gestes dans la cavité orale à une nouvelle configuration. Il compense les voyelles demandées par d'autres réalisations vocaliques, plus ou moins éloignées, en termes de « cibles » spatiotemporelles, mais qu'il peut réaliser plus aisément. Ainsi, ce patient produit des sons qui se rapprochent plus de [e] et [ʊ] pour la réalisation des voyelles [i] et [u], respectivement. En dépit des effets secondaires de fibroses musculaires par exemple, certains auteurs montrent que l'amélioration fonctionnelle peut se poursuivre après la radiothérapie (Furia *et al.*2001 ; Hsiao *et al.*2003). Le travail musculaire personnel à travers le temps et avec la rééducation orthophonique, permet une amélioration de la mobilité linguale et donne un peu de souplesse au lambeau initialement inerte. Cette amélioration est perçue dans les autres phases post-opératoires 2 et 3. Les triangles ont tendance à s'élargir et donc à tendre vers une normalisation des valeurs de F1 et F2 (*Cf.* figure 2).

(Birl) (Jct)

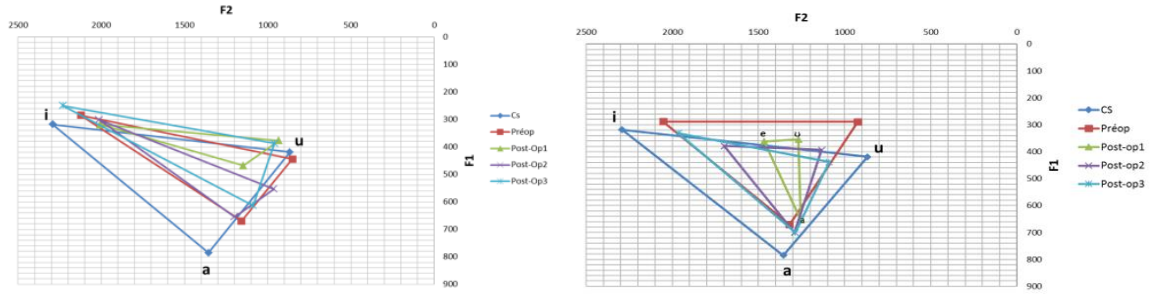



**Figure 2** : Valeurs formantiques de F1 et F2 des voyelles tenues [i,a,u] après une pelvi-glosso-mandibulectomie (Birl) à gauche et d'une glossectomie totale (Jct) à droite suivie d'une reconstruction

## 4.2 Aire de l'espace vocalique

Des analyses de variance (*ANOVA*) à deux facteurs ont été effectuées pour la variable aire des espaces vocaliques, afin de déterminer s'il existait des effets de *phases d'enregistrement (temps)*. Les deux facteurs *phase d'enregistrement* et *patient* (type d'exérèse) se sont révélés significatifs (*p<0.0001*) pour la variable aire de l'espace vocalique (*cf.* tableau 2, les résultats des analyses statistiques).

| Type d'exérèse et identification du patient/ Phase | Préop- Postop1 | Préop- Postop2 | Préop- Postop3 | Postop1-Postop2 | Post-op1- Postop3 | Post-op2- Postop3 |
|---|---|---|---|---|---|---|
| Gloss T + reconstruction (JCT) | *** | * | ns | ns | * | ns |
| Pelvi-glosso-Mandibul + reconstruction (BIRL) | ns | ns | ns | ns | ns | ns |
| Pelvi-gloss (PETR) | ns | * | ns | ns | ns | ns |
| Gloss P 1 (ZIM) | ns | ns | ns | ns | ns | ns |
| Gloss P 2 (SOM) | ns | ns | ns | ns | ns | ns |
| Gloss P 3 (Glad) | * | ns | ns | ns | ** | ns |
| Gloss P 4 (SIB) | ns | ns | ns | ns | ns | ns |
| Gloss P 5 (ANT) | * | ns | ns | ns | ns | ns |
| Gloss P 6 (ROJ) | * | ns | ns | ns | ns | ns |
| Hémi-gloss + reconstruction (HACH) | ns | ns | ns | ns | ns | ns |

**Tableau 2 :** Résultats des analyses statistiques de l'effet phases d'enregistrement sur l'aire de l'espace vocalique chez les 10 patients glossectomisés

Les résultats des analyses statistiques ont été obtenus en reliant les valeurs les plus extrêmes entre elles, en calculant à partir de la formule de Héron, l'aire de l'espace maximale de chaque répétition au sein de chaque phase, pour chaque patient sujet de notre étude.
Nous pouvons constater que l'aire de l'espace vocalique est significativement réduite (*p<0.0001*) pour le patient (Jct), entre les phases d'enregistrement préop et post-op1, avant de ré-augmenter (*p<0.05*) en phase post-op2, sans pour autant atteindre la valeur de référence de la phase préopératoire. Elle est de (1.26 kHz²), (0.43 kHz²), (0.71 kHz²), (0.99 kHz²) respectivement. Pour la patiente (Petr), l'aire de l'espace vocalique en post-op2 est également modifiée (*p<0.05*) de manière significative (1.57kHz²), par rapport à celle relevée en post-op1 (1.77kHz²). Nous avons également relevé que l'aire de l'espace vocalique est significativement réduite (*p<0.05*) en phase post-op1 pour les patients (Ant, Glad et Roj). En préop, elle est de (1.51 kHz², 1.78 kHz² et 1.68 kHz²), pour se réduire à (1.03 kHz² ,1.32 kHz² et 1.30 kHz²) en post-op1. Notons qu'à partir de la phase post-op 2, pour certains de nos patients et du post-op3 pour d'autres, l'aire de l'espace vocalique augmente pour atteindre les valeurs de départ relevées en préop, preuve d'un réajustement (*p=ns*). L'espace vocalique est géométriquement conventionnel (*p=ns*) pour 5 patients sur 6 patients ayant subi une glossectomie partielle, et les différences restent minimes entre les phases d'enregistrement. C'est aussi le cas pour la patiente (Hach) ayant subi une hémi-glossectomie, suivie d'une reconstruction avec un lambeau anté-brachial.

## 4.3 L'indice (PHi)

Les résultats de mesures de l'indice de dispersion de l'organisation du système vocalique montrent qu'en préop, les valeurs de l'indice (PHi) restent assez élevées, une indication d'une bonne organisation du système vocalique. En post-op1, suite à la chirurgie et la reconstruction pour certains de nos patients, ainsi que les effets secondaires des traitements complémentaires en post-op

2, nous avons remarqué une importante baisse des valeurs du (PHi), ce qui révèle une perturbation et une désorganisation du système vocalique.

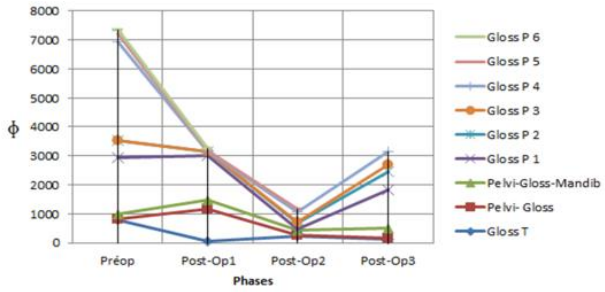



**Figure 3 :** Résultats des mesures de l'indice de dispersion de l'organisation du système vocalique PHi(ϕ) à travers les phases d'enregistrement chez les 10 patients glossectomisés

En post-op3, et après réhabilitation pour certains de nos patients, les valeurs du (PHi) augmentent proportionnellement selon le type d'exérèse et le traitement qu'a subi chaque patient (*cf.* figure 3). Ainsi, le patient tente d'adapter ses capacités gestuelles dans la cavité buccale à une nouvelle configuration, ce qui aboutit à une réorganisation du système vocalique. En effet, plus la valeur du (PHi) est élevée, plus la dispersion des entités vocalique dans le système vocalique reste contrainte et organisée. Plus la valeur du (PHi) est basse, plus les éléments vocaliques dans ce système se disperse, donnant un système globalement désorganisé.

## 5 Discussion et conclusion

Nous proposons de vérifier les hypothèses émises au départ, et de savoir si elles ont été confirmées ou infirmées. (1) Plus la taille de la tumeur est large plus les répercussions sur la production de la parole sont importantes étant donné que cela implique une reconstruction. Les valeurs formantiques, extraites à partir des voyelles soutenues, ont été effectivement remarquablement perturbées. (2) Les modifications manifestées au niveau des valeurs formantiques de F1 et F2 se répercutent forcément sur la taille et la forme de l'espace vocalique. Ce dernier est plus restreint en post-op 1 et 2 et a changé de configuration, principalement pour les patients qui ont subi des reconstructions linguales. Suite à la nouvelle modification anatomique, ces patients mettent plus de temps à réapproprier leur cavité buccale et adopter de nouvelles stratégies gestuelles dans cette configuration du tractus buccal, par rapport au patient ayant subi une glossectomie partielle. Ces résultats corroborent ceux attestés dans le littérature (Bressman *et al.* 2005, 2007 ; Acher *et al.*2014). (3) Plus l'indice de mesure de dispersion de l'organisation du système vocalique est élevé, plus l'organisation du système vocalique est adéquate, et *vice-versa.* (4) Enfin, la rééducation orthophonique et le facteur temps ont un effet positif sur tous les paramètres mesurés dans notre étude. Une amélioration dans la production de la parole chez les sujets atteints a été perçue dans les phases post-opératoires 2 et 3.

Ce travail nous a permis de vérifier la pertinence des mesures acoustiques pour évaluer les importantes perturbations que peut entrainer l'ablation de la langue sur la parole de certains de nos patients. Plus l'étendue de la lésion est large, plus les répercussions sont importantes. Cette étude a également permis d'observer les stratégies compensatoires que les patients sont capables de mettre en place seul ou à l'aide d'une rééducation orthophonique. Ces stratégies peuvent être soit conservatrice et qui consiste en une modification géométrique du tractus vocal par des synergies modifiées des différents articulateurs, ou bien innovatrices entrainant une nouvelle forme du tractus vocal, due à la reconstruction (Vaxelaire, 2007). Ainsi, les stratégies de compensation requièrent le recrutement d'autres structures saines (mandibule, lèvres), afin de pallier, à différents degrés,

l'affaiblissement ou la perte de mobilité linguale nécessaire pour atteindre les « cibles » articulatoires. Notons que la variabilité interlocuteur est toujours très importante dans les phases d'enregistrement post-opératoires. Il pourrait donc se révéler intéressant d'augmenter le nombre de patients afin de pouvoir classer les patients dans des sous-groupes « homogènes », suivant les conséquences de la chirurgie sur la parole des patients. Sur le plan perceptif, il serait intéressant d'effectuer des tests de perception afin de comparer l'altération des voyelles perçue par un jury d'écoute et le spectre altéré, puis amélioré de ces mêmes voyelles. D'autres données recueillies auprès des patients glossectomisés (logatomes, questionnaire d'auto-évaluation du ressenti (SHI)) sont en cours d'exploitation.

## Références


-Acher A., Perrier P., Savariaux C., & Fougeron C. (2014). Speech production after glossectomy: Methodological aspects. *Clin Linguist Phon 28* (4), 241-256.

-Bressmann T., Sader R., Whitehill T. L., & Samman N. (2004). Consonant intelligibility and tongue motility in patients with partial glossectomy. *Journal of Oral and Maxillofacial Surgery, 62,* 298-303.

-Buchaillard S., Brix M., Perrier P, & Payan Y (2007). Simulations of the consequences of tongue surgery on tongue mobility: Implications for speech production in post-surgery conditions. *International Journal of Meical Robotics and Computer Assisted Surgery*, *3*(3), 252-261.

-Crevier-Buchman L., Smadja M., Tessier C., Menard M., Brasnu D. (2007). Evaluation de la qualité de vie après glossectomie partielle. Revue Française d'ORL, 288-301.

-Diz Dios P., Fernandez Feijoo J., Castro Ferreiro M. (1994). Functional Consequences of Partial Glossectomy. *Journal of Oral and Maxillo-facial Surgery, 52*(1), 12-14.

-Furia C., Kowalski L., Latorre M., Angelis E., Martins N., Barros A., Ribeiro K. (2001). Speech intelligibility after glossectomy and speech rehabilitation, *Archives of Otolaryngology, Head and Neck Surgery*, Chicago.127, (7), 877-883

-Heller K S., Levy J & Sciubba, J. J (1991). Speech patterns following partial glossectomy for small tumors of the tongue. *Head Neck*, 13, 4, 340–343.

-Hsiao H.-T., Leu Y.-S., Lin C.-C. (2003) Tongue reconstruction with free radial forearm flap after hemiglossectomy: a functional assessment. *Journal of Reconstructive Microsurgery*. New York, Theme Medical Publishers, 19, (3), 137-142

-Huet K & Harmegnies B. (2000). Contribution à la quantification du degré d'organisation des systèmes vocaliques. *In Actes des JEP'2000*, 225-228.

-Korpijaakko-Huuhka A.-M., Söderholm A.-L. (1998). Long-lasting speech and oral motor deficiencies following oral cancer surgery: a retrospective study. *Logopedics Phoniatrics Vocology*, 24 (3), 97-106.

-Meley M & Barthelmé E. (1987). Les cancers de la cavité buccale et de l'oropharynx. France. Masson. 164

-Savariaux C., Perrier P., Pape D., Lebeau J (2001). Speech production after glossectomy and reconstructive lingual surgery: a longitudinal study. *In Proceedings of the 2nd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, Firenze, Italy.

-Shin Y. S., Koh Y. W., Kim S.-H., Jeong J. H., Ahn S., Hong H. J., & Choi E. C. (2012). Radiotherapy deteriorates postoperative functional outcome after partial glossectomy with free flap reconstruction. *J Oral Maxillofac Surg 70* (1), 216–220.

-Vaxelaire B. (2007) La résistivité spatio-temporelle de gestes linguistiques. Ou comment perturber le linguistique en augmentant la vitesse d'élocution. *In* « Perturbations et réajustements : langue et langage. » B. Vaxelaire R. Sock G. Kleiber F. Marsac (Eds.), Publications de l'Université Marc Bloch de Strasbourg. 179-199.

# Efforts de production de parole chez les personnes qui bégaient

Maëva Garnier, Anaïs DaFonseca, Christophe Savariaux, Thibault Cattelain,
Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-lab, 38000 Grenoble, France
* Institute of Engineering Univ. Grenoble Alpes
maeva.garnier@gipsa-lab.grenoble-inp.fr

## RÉSUMÉ

Cette étude vise à objective et quantifier les tensions orofaciales visibles et communément décrites chez des personnes qui bégaient (PQB), en comparant les efforts aréeodynamiques, laryngés et articulatoires lors de la production de consonnes occlusives entre des PQB et des personnes normofluentes (PNF). Dans cette étude préliminaire, 4 adultes qui bégaient et 4 adultes normofluents ont été enregistrés. Ils devaient produire des mots commençant par des consonnes occlusives labiales dans une tâche de lecture et dans une tâche semi-spontanée. Contrairement à nos attentes, les 4 PQB ont montré une force interlabiale (mesurée à partir d'un capteur de pression) et une pression intra-orale réduites par rapport aux 4 PNF. Aucune différence significative n'a été observée au niveau de ces paramètres entre les syllabes bégayées v. perceptivement fluentes des PQB. Cependant, les différences de pression intra-orale observes entre les PQB et le PNF sont renforcées sur des syllabes complexes (commençant par un cluster consonantique) et pour la tâche semi-spontanée.

## ABSTRACT

**Speech production efforts in people who stutter.**
This study aims at objecting and quantifying the visible oro-facial tensions that are commonly described in people who stutter (PWS), by comparing the aerodynamic, laryngeal and articulatory efforts of speech production between PWS and normo-fluent people (NFP). 4 adult PWS and 4 NFP were recorded, while producing words beginning with labial stop consonants in a reading task and a semi-spontaneous task. Contrary to our expectations, PWS demonstrated less intra-oral pressure and interlip force (measured with a force sensor) than NFP (with comparable syllable intensity, though). No significant difference in intra-oral pressure and interlip force was observed between the disfluent vs. perceptually fluent syllables produced by PWS. However, the differences in intra-oral pressure observed between PWS and NFP were greater on complex syllables (beginning with a consonant cluster) and for the semi-spontaneous task. The electroglottographic data is currently being analyzed and will be presented at the conference.

MOTS-CLES : Bégaiement, efforts de production, consonnes occlusives, pression intra-orale, force interlabiale.
KEYWORDS: Stuttering, production efforts, stop consonants, intra-oral pressure, interlip force.

# 1 Introduction

Le bégaiement est un trouble neuro-moteur se traduisant par des difficultés et des tensions lors de la production de parole, allant jusqu'à différents types de disfluences particulièrement audibles : des répétitions, des prolongations ou des blocages de sons.


10.21437/JEP.2018-10

La question se pose encore de savoir si ce trouble est intermittent, n'existant qu'au moment des difluences perceptibles, provoqué par des facteurs extérieurs (stress, fatigue, charge cognitive, …) (Caruso et al., 1994), ou bien si ce trouble est permanent, se traduisant à tout moment par des atypicités motrices, mais dont les conséquences acoustiques sont plus ou moins perceptibles et se répartissent sur un continuum de fluence (Peters et al. 2000).

La question se pose également de savoir quel(s) geste(s) sont altérés par ce trouble.
La plupart des études a porté sur la description de gestes laryngés atypiques (Montfrais-Pfauwadel 2005) et d'une activité musculaire laryngée excessive (Freeman & Ushijima, 1978) au moment des disfluences. On dispose de moins d'information sur ce qui se passe au niveau articulatoire et respiratoire au moment des bégayages. Il a toutefois été montré que les disfluences perceptibles étaient généralement précédées de mouvements de la mâchoire (d'ouverture ou de fermeture) dont la vitesse était particulièrement élevée (Hutchinson and Watkin, 1976), d'anomalies au niveau des mouvements des lèvres et de la langue (Didirkova, 2016), associés à une augmentation plus rapide que la moyenne de la pression intra-orale pendant une occlusion (Hutchinson and Navarre, 1977).
Par ailleurs, il semblerait que les gestes de production de parole des personnes qui bégaient continuent de montrer des particularités, même en dehors des instants de disfluences audibles. Ainsi, différents auteurs ont montré chez les PQB une hyper-excitabilité musculaire latente au niveau laryngé (Freeman & Ushijima, 1978) et des difficultés à coordonner rapidement leurs mouvements laryngés et respiratoires pour initier ou terminer la vibration des plis vocaux (Adams 1974). Dans leur parole en apparence fluente, les PQB montrent également un débit oral expiré et une pression intra-orale significativement plus faibles que les PNF(Hutchinson et Navarre, 1977) et des défaillances du contrôle de la pression sous-glottique (Peters & Boves, 1988). Des résultats beaucoup plus variables, voire contradictoires, sont reportés au niveau articulatoire : Van Lieshout et al. (1993) retrouvent, comme au niveau laryngé, des activités musculaires plus élevées au niveau de la lèvre inférieure des PQB, durant ou précédant la parole. Mais d'autres études n'observent aucune différence significative d'activité musculaire orofaciale entre des PQB et des PNF. Certains auteurs (Smith, 1989; Denny and Smith, 1992 ; De Felicio et al. 2007) observent même, au contraire, une activité musculaire inférieure à la moyenne au niveau de la lèvre supérieure. Une réduction vocalique est classiquement observée chez les personnes qui bégaient (Hirsch, 2007; Blomgren et al., 1998; Klich and May, 1982), supportant l'idée de gestes articulatoires moins amples. Pour autant, seul Zimmermann (1980) et McClean & Runyan (2000) observent effectivement des mouvements moins amples et moins rapides de la lèvre inférieure et de la mâchoire. D'autres études ne trouvent aucune différence significative d'amplitude, de vitesse et/ou de durée des mouvements des lèvres ou de la mâchoire dans la parole perceptivement fluente de PQB, comparées à des PNF (McClean et al., 1990; Loucks et al., 2007). Enfin d'autres études encore rapportent de plus grandes amplitudes, de plus grands pics de vitesse, et de plus grandes durées des mouvements de la lèvre supérieure et de la langue chez les personnes qui bégaient (Namasivayam & van Lieshout, 2008 ; Zmarich, 1994; 2001 ; Max et al., 2003; McClean and Runyan, 2000). A notre connaissance, aucune information n'a été apportée quant à la force des mouvements articulatoires d'occlusion.

Notre objectif est ici de décrire et comprendre ce que se passe au niveau du geste de production de parole lorsque la personne bégaye mais aussi le reste du temps lorsque sa parole semble en apparence fluente. En nous intéressant aux niveaux respiratoire et articulatoire, nous nous proposons

- d'observer les manifestations d'effort et les indices de difficultés de coordination entre les différents niveaux observés
- d'objectiver les tensions et les difficultés de production ressenties par les personnes qui bégaient en mesurant quantitativement la force articulatoire interlabiale, la pression intra-orale et l'intensité acoustique de consonnes occlusives bilabiales (/p/, /b/, /m/).

- d'apporter des éléments de réponse quant à l'aspect local ou global du bégaiement et quant à sa nature permanente ou intermittente, en comparant les efforts respiratoires, laryngés et articulatoires entre des locuteurs adultes qui bégaient et des locuteurs sans troubles de la parole, et entre les productions fluentes et disfluentes de personnes qui bégaient.

Nous formulons et testons ici 4 hypothèses :

H1 : Les personnes qui bégayent produiraient des consonnes occlusives bilabiales avec plus d'effort articulatoire et respiratoire que les personnes sans trouble de la parole

H2 : Des différences significatives de production de la parole seraient particulièrement observées pendant les bégayages de personnes qui bégaient, comparés à la parole de personnes normofluentes, mais également de façon moindre dans les autres segments de parole perceptivement fluents des personnes qui bégaient

H3 : Ces différences significatives de production de la parole entre personnes normofluentes et personnes qui bégaient seraient amplifiées par la complexité phonologique.

H4 : Ces différences significatives de production de la parole entre personnes normofluentes et personnes qui bégaient seraient observées uniquement, ou davantage, en tâche semi-spontanée par rapport à une tâche de lecture.

# 2 Matériel et méthodes

Quatre adultes qui bégaient (PQB) (2 hommes, 2 femmes) et quatre adultes normofluents (PNF) (3 hommes, 1 femme), âgés de 20 à 46 ans, tous francophones, ont participé à cette étude préliminaire, réalisée à l'occasion du mémoire d'orthophonie d'Anaïs Da Fonseca (2016). Les quatre PQB présentaient un bégaiement développemental et avaient suivi des rééducations orthophoniques par le passé (de 3 à 5 ans selon les sujets, avec des méthodes variables dont ERASM). Plus aucun participant ne suivait de rééducation au moment de l'expérience.

Les participants ont été enregistrés dans deux conditions expérimentales : 1) une condition non-interactive de lecture de listes de mots le plus rapidement possible et 2) une condition de production semi-spontanée de ces mêmes mots : l'expérimentateur, assis en face du participant, retournait sur la table des séries de 3 cartes représentant les mots cibles. Le participant devait rapidement inventer une phrase à partir de ces 3 mots. Chaque série était répétée 5 fois dans un ordre pseudo-aléatoire afin de composer des phrases différentes pour chaque série.

Le tableau 1 récapitule les 30 mots cibles de l'expérience. 18 d'entre eux ont été sélectionnés pour être constitués de 3 syllabes (CVCVCV), avec /p/, /b/ ou /m/ en consonne initiale[1], suivie de la voyelle /a/, /i/ ou /o/, et de genre masculin de façon à ce que la consonne occlusive soit précédée de la voyelle /e/. Les 12 mots restants ont été sélectionnés sur le modèle, mais débutant par les clusters consonantiques /pR/ et /bR/ de façon à examiner l'effet de la complexité articulatoire.

Quatre signaux ont été acquis simultanément :

- La pression intra-orale a été enregistrée à l'aide d'un fin tube capillaire maintenu dans la cavité buccale, relié à un pneumotachographe (station d'acquisition de données physiologiques EVA).
- Le signal audio de parole a été enregistré à l'aide d'un microphone de pression (Bruël and Kjær 4944-A) placé à 30cm des lèvres. Le niveau d'intensité acoustique était calibré à l'aide d'un amplificateur de mesure (Nexus, Bruel&Kjaer) délivrant un signal interne de référence à 1kHz.
- Le signal électroglottographique a été enregistré à l'aide d'un électroglottographe (EG2 Glottal Enterprise).

1 Les personnes qui bégaient rencontrent des difficultés particulières sur les consonnes occlusives bilabiales (Didirkova 2016)

- La force de compression interlabiale a été mesurée à l'aide d'un capteur de pression à jauge de contrainte, collé sur la lèvre inférieure (Jeannin et al. 2008 ; Garnier et al. 2014).
- Ces trois derniers signaux ont été numérisés à 44100 Hz sur une même carte d'aquisition (Biopac) et post-synchronisés avec le signal de Pio acquis sur la station EVA.

| | **/p/** | **/pʀ/** | **/b/** | **/bʀ/** | **/m/** |
|---|---|---|---|---|---|
| **/a/** | Paradis | Praticien | Bananier | Brasero | Macaron |
| | Panama | Praliné | Baluchon | Braconnier | Maquillage |
| **/i/** | Pissenlit | Prisonnier | Bikini | Bricoleur | Mirabelle |
| | Piranha | Professeur | Bijoutier | Britannique | Minibus |
| **/o/** | Potager | Promoteur | Boléro | Brocoli | Mocassin |
| | Policier | Privation | Bolognaise | Brocanteur | Mobilier |

TABLE 1 : Liste des 30 mots cibles visant à examiner la production de consonnes occlusives dans différents contextes vocaliques et dans des clusters consonantiques.

Les données ont été étiquetées manuellement sous Praat en repérant :
- le début et la fin des mots-cibles
- le début et la fin des syllabes initiales (CV ou CCV)

A l'aide de scripts développés sous Matlab, différents descripteurs ont ensuite été extraits des signaux enregistrés sur ces différents intervalles de temps :
- le maximum de pression intra-orale pendant la phase d'occlusion de la consonne (Piomax).
- le maximum de force interlabiale pendant la phase d'occlusion de la consonne (Fmax).
- l'intensité acoustique de la syllabe (Imax).

Les analyses statistiques ont été réalisées avec le logiciel R sur les syllabes fluentes des PQB et PNF. Pour chaque descripteur, nous avons modélisé les données à l'aide de modèles mixtes, incluant un effet aléatoire « Locuteur » et trois facteurs à effet fixe :
- Catégorie de sujet (2 niveaux : PQB vs. PNF). Les sujets n'étaient pas appariés.
- Tâche (2 niveaux : lecture vs. semi-spontanée)
- Consonne ou cluster de consonnes en début de mot (5 niveaux : /p/, /b/, /m/, /pr/, /br/)

Nous avions choisi 3 contextes vocaliques /a/, /i/ et /o/ non pas pour étudier l'influence du contexte vocalique mais plutôt pour que nos résultats ne soient pas spécifiques à une voyelle donnée. C'est pourquoi, en l'absence d'hypothèse sur l'effet du contexte vocalique, nous ne l'avons pas considéré dans les facteurs du modèle statistique.
Suivant la même procédure que celle développée avec des experts statisticiens, appliquée à l'analyse de plusieurs jeux de données précédents (Bourne et al. 2016), nous avons commencé par chercher à simplifier chaque modèle en excluant toute interaction non significative entre les facteurs à effet fixe. Pour cela, nous avons utilisé la fonction step sous R. Une fois le modèle simplifié, nous avons vérifié sa validité en examinant ses résidus. Nous avons ensuite réalisé des tests de modèle emboités (ou Likelihood Ratio Test) pour tester la significativité de chaque facteur ou de leurs interactions. Enfin, nous avons réalisé des tests posthoc pour examiner le contraste plus spécifique entre certaines conditions, en appliquant des corrections de Bonferroni pour les comparaisons multiples.
Du fait de l'effectif réduit des syllabes bégayées, la significativité des différences entre syllabes bégayées et perceptivement fluentes des PQB n'a pu être testée statistiquement.

# 3 Résultats

La Figure 1 représente le taux de bégayages pour chacune des 4 personnes qui bégaient, en fonction de la syllabe initiale des mots et en fonction de la tâche de parole. En revanche, le niveau de bégayage est assez comparable entre les 4 individus de notre expérience, en moyenne de 2.72%.
Contrairement à nos attentes, les bégayages ne sont pas significativement plus fréquents sur les syllabes phonologiquement plus complexes. Les bégayages les plus fréquents sont effectivement observés sur les mots-cibles commençant par /pR/, mais cette tendance ne s'étend pas aux mots-cibles commençant par /bR/. Conformément à nos attentes, les bégayages tendent à être plus fréquents sur la tâche semi-spontanée, comparée à celle de lecture des mots-cibles.

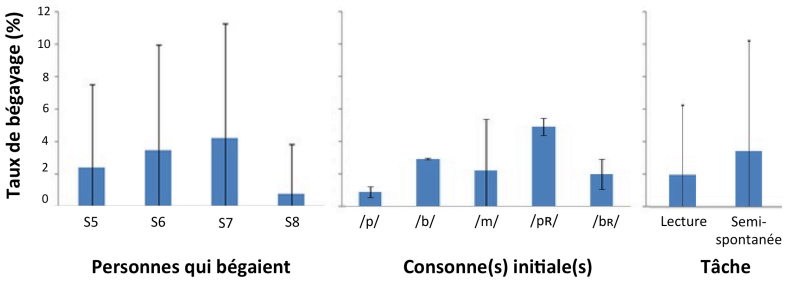



FIGURE 1 : Taux de bégayages (moyenne et écart-type) en fonction des participants, de la syllabe initiale et de la tâche.

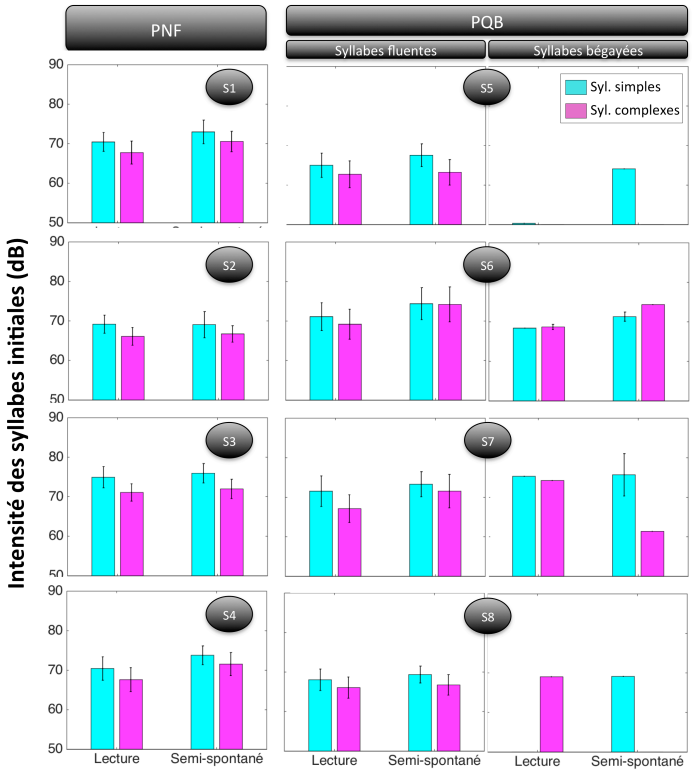



FIGURE 2 : Intensité acoustique (moyennes et écart-types) des syllabes initiales chez les 4 personnes qui bégaient (PQB), pour les syllabes perceptivement fluentes ou bégayées et des 4 personnes normo-fluentes (PNF), en fonction de la tâche de l'expérience (lecture ou semi-spontanée) et de la complexité de la syllabe.

La Figure 2 synthétise les résultats obtenus quant à l'intensité acoustique des syllabes initiales. L'analyse statistique a montré que le modèle Intensité ~ Tache*TypeSyllabe expliquait le mieux la

variance de l'intensité vocale. Le facteur Groupe n'apparaît pas dans ce modèle simplifié, traduisant le fait que qu'il n'a pas d'effet significatif sur l'intensité des syllabes produites (df=1, LRatio=1.1, p>0.2). En d'autres termes, les 4 personnes qui bégaient de notre expérience parlent avec une intensité comparable aux 4 personnes normo-fluentes.

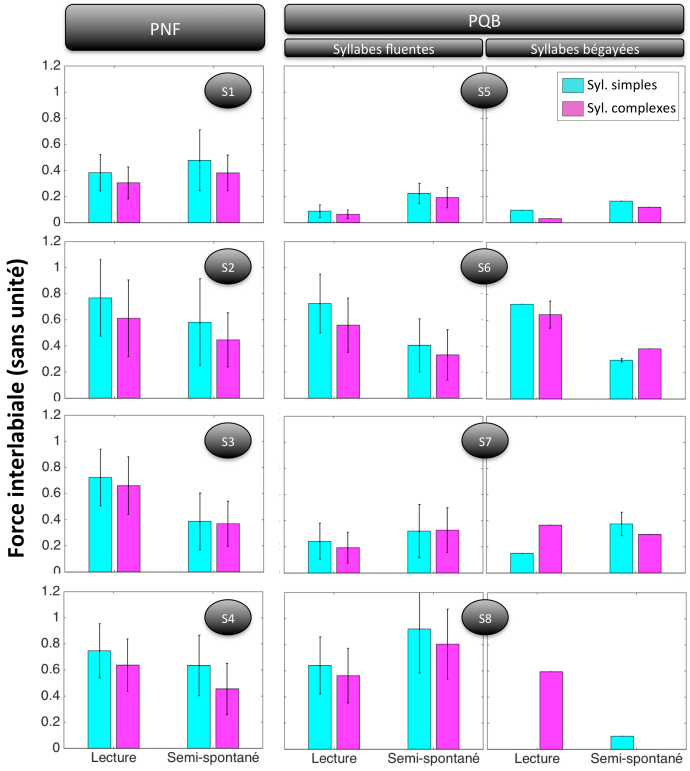



FIGURE 3: Force interlabiale (moyennes et écart-types) chez les 4 personnes qui bégaient (PQB), pour les syllabes perceptivement fluentes ou bégayées des 4 personnes normo-fluentes (PNF), en fonction de la tâche de l'expérience (lecture ou semi-spontanée) et de la complexité de la syllabe

La Figure 3 synthétise les résultats obtenus quant à la force articulatoire interlabiale mesurée sur les consonnes occlusives initiales des mots-cibles. L'analyse statistique a montré que le modèle Force ~ Groupe*Tâche + Groupe*TypeSyllabe expliquait le mieux la variance de la force interlabiale.

Il existe un effet d'interaction significatif entre le Groupe et la Tâche (df=1, Lratio=63.02, p<.0001 ***), traduisant le fait qu'en moyenne, les 4 personnes qui bégaient produisent des syllabes initiales avec une force articulatoire plus faible que les 4 personnes normofluentes en tâche de lecture (-0.22 (s.u), p=0.015, *), mais comparable lors de la tâche semi-spontanée (-0.03, p>0.9).

Par ailleurs, il existe également un effet d'interaction significatif entre le Groupe et le Type de syllabe initiale (df=4, Lratio=21.28,p<.0001 ***), mais celui-ci ne provient pas du fait que, en moyenne, la différence de force interlabiale entre les 4 PQB et les 4 PNF soit particulièrement accentuée ou diminuée par la complexité phonologique (Δ=0.03, p>0.3). Individuellement, on observe que tous les 4 PNF et 2 des PQB diminuent leur force articulatoire sur des syllabes complexes (commençant par /pR/, /bR/), comparée à des syllabes simples (commençant par /p/, /b/, /m/). Les deux autres PQB (S5, S7) ne varient pas significativement leur force interlabiale avec la complexité phonologique.

Enfin, la comparaison des syllabes bégayées et perceptivement fluentes produites par les 4 PQB ne montre pas de variation notable du niveau de force interlabiale (cf. Figure 3).

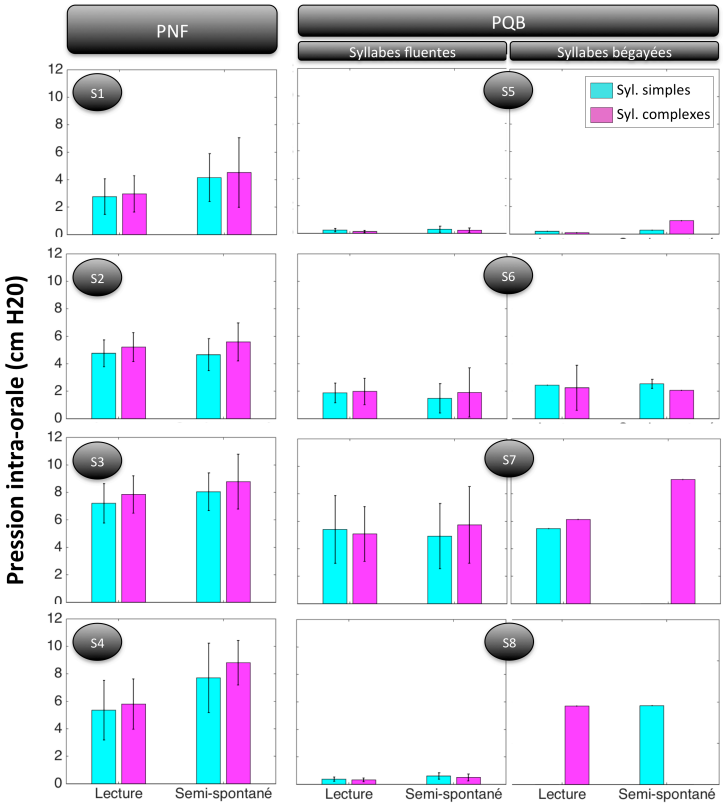



FIGURE 4: Pression intraorale chez les 4 personnes qui bégaient (PQB), pour les syllabes perceptivement fluentes ou bégayées des 4 personnes normo-fluentes (PNF), en fonction de la tâche de l'expérience (lecture ou semi-spontanée) et de la complexité de la syllabe

La Figure 4 synthétise les résultats obtenus quant à la pression intra-orale mesurée sur les consonnes occlusives initiales orales des mots-cibles (/p/, /b/) (la pression intra-orale est proche de zéro pour une occlusive nasale /m/, du fait que l'air continue de s'écouler par le nez).
L'analyse statistique a montré que le modèle global Pio ~ Groupe*Tâche*TypeSyllabe expliquait le mieux la variance de la pression intra-orale.
Malgré, l'interaction significative observée entre les trois facteurs (df=3, Lratio= 11.49, p=0.009 (**)), une différence très claire se dégage entre les 4 PQB et les 4 PNF de notre expérience, avec des niveaux de Pio significativement plus faibles chez PQB comparés aux PNF (-3.9 cm H2O en moyenne, p=0.003**).
Nous remarquons que cet écart de Pio entre les PQB et PNF est significativement plus marqué en tâche semi-spontanée qu'en tâche de lecture. ($\Delta$=+1.2 cm H20, p<.001 ***). En effet, on observe individuellement que les 4 locuteurs bègues, ainsi que le locuteur normo-fluent S2 ne varient pas significativement leur niveau de Pio entre les 2 tâches, tandis que les 3 autres locuteurs normofluents (S1, S3 et S4) augmentent leur Pio entre la tâche de lecture et la tâche de parole semi-spontanée (cf. Figure 4).

On note également que l'écart de Pio entre les PQB et PNF est légèrement plus marqué pour les syllabes complexes que les syllabes simples. ($\Delta$=+0.5 cm H20, p=0.01 *). Individuellement, on observe en effet que tous les locuteurs normofluents augmentent leur Pio sur des syllabes complexes (commençant par /pR/, /bR/), comparées à des syllabes simples (commençant par /p/, /b/, /m/), tandis qu'aucun locuteur bègue ne montre de variation du niveau de Pio avec la complexité phonologique.
Enfin, la comparaison des syllabes bégayées et perceptivement fluentes produites par les 4 PQB ne montre toujours pas de variation notable du niveau de pression intra-orale pour 3 des locuteurs (S5, S6, S7). Seul le locuteur S8 montre une grande augmentation de pression intra-orale lors des bégayages (cf. Figure 4).

# 4 Discussion et conclusion

Contrairement à notre hypothèse, les 4 PQB de cette étude ont montré moins de Pio et de force interlabiale que les PNF (malgré une intensité vocale comparable). Aucune différence significative de Pio et de force interlabiale n'a été observé entre les syllabes bégayées vs. perceptivement fluentes des PQB. Les différences de Pio observées entre nos 4 PQB et nos 4 PNF étaient plus importantes sur des syllabes complexes et pour une tâche semi-spontanée, ce qui n'était pas le cas pour la force inter-labiale.

Bien que contraires à nos attentes, les moindres forces aérodynamiques et articulatoires observées chez les PQB sont cohérentes avec d'autres observations reportées par de précédents études sur le bégaiement, en particulier une pression intra-orale (Pio) plus faible chez des PQB (Hutchinson & Navarre, 1977), une amplitude et des vitesses réduites des gestes articulatoires (McClean & Runyan, 2000; Zimmermann, 1980; Zmarich, 2001), une réduction vocalique…) (Blomgren, Robb, & Chen, 1998; Hirsch, 2007; Klich & May, 1982).

Il est possible que les niveaux plus faibles de force interlabiale et de pression intra-orale mesurés chez les 4 PQB de cette étude reflètent effectivement un certain niveau de décontraction, correspondant à la mise en œuvre de techniques d'aide à la fluence du type ERASM, à laquelle nos 4 PQB, comme la plupart des adultes
Bègues, sont formés à un moment ou un autre de leur parcours thérapeutique. Une autre explication serait que les tensions faciales visibles chez les personnes qui bégaient et leurs niveaux d'activité musculaire plus élevés ne soient pas nécessairement reliés à une amplification des gestes articulatoires ou respiratoires. Au contraire, certains auteurs ont pu observer une co-contraction de muscles antagonistes au niveau laryngé (Freeman & Ushijima, 1978), pouvant expliquer, selon ces auteurs, la lenteur des mouvements d'adduction/abduction laryngée. Il est donc tout à fait envisageable qu'il en soit de même au niveau articulatoire et respiratoire, et que les tensions faciales visibles des personnes qui bégaient et leurs niveaux d'activité musculaire plus élevés témoignent en fait de crispations et de co-contraction de groupes de muscles antagonistes, ce qui aurait plutôt pour effet de réduire, voire de bloquer les mouvements, plutôt que d'amplifier leur force comme nous le pensions initialement.

# Remerciements

Nous remercions les 8 participants à cette expérience. Cette recherche est financée par l'Agence Nationale de la Recherche (Projet StopNCo : Effort et coordination dans la production des consonnes occlusives ; ANR-14-CE30-0017; Maëva Garnier).

# Réferences


ADAMS, M. R. (1974). A physiologic and aerodynamic interpretation of fluent and stuttered speech. Journal of Fluency Disorders, 1(1), 35-47.

BOURNE T., GARNIER M. ET SAMSON A. (2016). Physiological and acoustic characteristics of the male music theatre voice. The Journal of the Acoustical Society of America, 140(1), 610-621

BLOMGREN, M., ROBB, M., & CHEN, Y. (1998). A note on vowel centralization in stuttering and nonstuttering individuals. Journal of Speech, Language, and Hearing Research, 41(5), 1042-1051.

CARUSO, A. J., CHODZKO-ZAJKO, W., BIDINGER, D., & SOMMERS, R. (1994). Adults Who Stutter: Responses to Cognitive Stress. Journal of Speech Language and Hearing Research, 37(4), 746.

DA FONSECA A. (2016). Effort articulatoire et respiratoire : étude de la parole bègue. Mémoire d'orthophonie, Université de Franche Comté.

DE FELÍCIO, C., FREITAS, R., VITTI, M., & REGALO, S. (2007). Comparison of upper and lower lip muscle activity between stutterers and fluent speakers. International Journal of Pediatric Otorhinolaryngology, 71(8), 1187-1192.

DENNY, M., & SMITH, A. (1992). Gradations in a Pattern of Neuromuscular Activity Associated With Stuttering. Journal of Speech Language and Hearing Research, 35(6), 1216.

DIDIRKOVA, I. (2016) Parole, langues et disfluences : une étude linguistique et phonétique du bégaiement. Thèse de doctorat de l'Université Paul Valéry - Montpellier III.

FREEMAN, F., & USHIJIMA, T. (1978). Laryngeal muscle activity during stuttering. Journal of Speech, Language, and Hearing Research, Vol. 21, 538-562.

GARNIER, M., BOUHAKE, S., ET JEANNIN, C. (2014) Efforts and coordination in the production of bilabial consonants. In 10th International Seminar on Speech Production. pp 138-141.

HIRSCH, F. (2007). Le bégaiement: perturbation de l'organisation temporelle de la parole et conséquences spectrales. Thèse de doctorat de l'Université de Strasbourg 2.

HUTCHINSON, J., & NAVARRE, B. (1977). The effect of metronome pacing on selected aerodynamic patterns of stuttered speech: Some preliminary observations and interpretations. Journal of Fluency Disorders, 2(3), 189-204.

HUTCHINSON, J., & WATKIN, K. (1976). Jaw mechanics during release of the stuttering moment : Some initial observations and interpretations. Journal of Communication Disorders, 9(4), 269- 279.

JEANNIN, C., PERRIER, P., PAYAN, Y., GROSGOGEAT, B., DITTMAR, A., & GÉHIN, C. (2009). PRESLA: An original device to measure the mechanical interaction between tongue and teeth or palate during speech production. In Proceedings of International Seminar on Speech Production.

KLICH, R., & MAY, G. (1982). Spectrographic study of vowels in stutterers' fluent speech. Journal of Speech, Language, and Hearing Research, 25(3), 364-370.

VAN LIESHOUT, P., PETERS, H., STARKWEATHER, C., & HULSTIJN, W. (1993). Physiological Differences Between Stutterers and Nonstutterers in Perceptually Fluent Speech: EMG Amplitude and Duration. Journal of Speech Language and Hearing Research, 36(1), 55.

LOUCKS, T. M. J., DE NIL, L. F., & SASISEKARAN, J. (2007). Jaw-phonatory coordination in chronic developmental stuttering. Journal of Communication Disorders, 40(3), 257-272.

MAX, L., CARUSO, A., & GRACCO, V. (2003). Kinematic Analyses of Speech, Orofacial Nonspeech, and Finger Movements in Stuttering and Nonstuttering Adults. Journal of Speech Language and Hearing Research, 46(1), 215.

MCCLEAN, M., KROLL, R., & LOFTUS, N. (1990). Kinematic Analysis of Lip Closure in Stutterers' Fluent Speech. Journal of Speech Language and Hearing Research, 33(4), 755.

MCCLEAN, M., & RUNYAN, C. (2000). Variations in the relative speeds of orofacial structures with stuttering severity. Journal of Speech, Language, and Hearing Research, 43(6), 1524-1531.

MCCLEAN, M., GOLDSMITH, H., & CERF, A. (1984). Lower-Lip EMG and Displacement During Bilabial Disfluencies in Adult Stutterers. Journal of Speech Language and Hearing Research, 27(3), 342.

MONFRAIS-PFAUWADEL, M.-C., TROMELIN, O., MOUGIN, A.-L., & ORMEZZANO, Y. (2005). Utilisation des explorations multimédia synchrones dans l'objectivation des événements laryngés lors des bégayages. In Revue de laryngologie, d'otologie et de rhinologie (Vol. 126, p. 341-345).

NAMASIVAYAM, A. K., & VAN LIESHOUT, P. (2008). Investigating speech motor practice and learning in people who stutter. Journal of Fluency Disorders, 33(1), 32-51.

PETERS, H. & BOVES, L. (1988). Coordination of aerodynamic and phonatory processes in fluent speech utterances of stutterers. Journal of Speech, Language, and Hearing Research, 31(3), 352-361.

PETERS, H. M., HULSTIJN, W., & VAN LIESHOUT, P. H. (2000). Recent developments in speech motor research into stuttering. Folia Phoniatrica et Logopaedica, 52(1-3), 103-119.

SMITH, A. (1989). Neural Drive to Muscles in Stuttering. Journal of Speech Language and Hearing Research, 32(2), 252.

ZIMMERMANN, G. (1980). Articulatory Dynamics of Fluent Utterances of Stutterers and Nonstutterers. Journal of Speech Language and Hearing Research, 23(1), 95.

ZMARICH. (1994). Articulatory kinematics of lips and jaw in reiterant/pa/and/ba/sequences in Italian stutterers. Journal of Fluency Disorders.

ZMARICH, C. (2001). What Phonetics has to say about stuttering. Lecture held at the Socrates European Intensive Programme for Speech and Language Therapy, Diploma Universitario di Logopedia, Collegio Mazza, Padova, 22, 2000.

# Segmentation et Regroupement en Locuteurs: comment évaluer les corrections humaines

Broux Pierre-Alexandre[1, 2] Doukhan David[2] Petitrenaud Simon[1]
Meignier Sylvain[1] Carrive Jean[2]
(1) Laboratoire informatique de l'université du Maine (LIUM - EA 4023), Avenue Olivier Messiaen, F-72085 Le Mans, France
(2) Institut national de l'audiovisuel (Ina), 18 Avenue des Frères Lumière, 94360 Bry-sur-Marne, France
pabroux@ina.fr, ddoukhan@ina.fr, simon.petit-renaud@univ-lemans.fr, sylvain.meignier@univ-lemans.fr, jcarrive@ina.fr

RÉSUMÉ
Dans cet article, nous présentons un simulateur dédié à l'évaluation des corrections humaines sur la tâche de Segmentation et Regroupement en Locuteurs (SRL). Nous proposons quatre actions élémentaires afin de corriger une SRL et un automate pour simuler la séquence de corrections. Une mesure est proposée pour évaluer le coût de correction. Le simulateur est évalué en utilisant des émissions françaises d'information tirées du corpus REPERE.

ABSTRACT
**Computer-assisted speaker diarization : how to evaluate human corrections**

In this paper, we present a framework to evaluate the human correction of a speaker diarization. We propose four elementary actions to correct the diarization and an automaton to simulate the correction sequence. A metric is described to evaluate the correction cost. The framework is evaluated using French broadcast news drawn from the REPERE corpus.

MOTS-CLÉS : Segmentation et Regroupement en Locuteurs, annotation, Interactions Homme-Machine (IHM), évaluation.

KEYWORDS: Speaker diarization, annotation, Human-Computer Interaction (HCI), evaluation.

## 1 Introduction

Le travail présenté dans cet article a été réalisé pour répondre à certains objectifs de l'Institut national de l'audiovisuel (Ina). L'Ina est une institution publique en charge de la préservation et de la valorisation du patrimoine audiovisuel français. La valorisation repose en partie sur l'annotation de collections de documents audiovisuels. L'annotation consiste à enrichir les documents avec des titres, des résumés, des mots-clés ou les noms des participants pour répondre aux requêtes des clients et des usagers de l'Ina, qu'ils soient journalistes, producteurs, réalisateurs ou chercheurs. Cependant, en raison du nombre croissant de documents et du nombre limité de documentalistes, beaucoup de documents restent peu ou pas documentés. Les informations fournies par la documentation varient grandement selon le type d'archives : les émissions d'information (journaux télévisés et magazines d'actualité) sont habituellement finement documentées, tandis que d'autres programmes tels que les jeux, les documentaires, les émissions de variété ou de téléréalité le sont plus sommairement.


10.21437/JEP.2018-11

Une des solutions pour faciliter l'annotation et améliorer l'accès aux documents est d'utiliser les technologies de reconnaissance automatique de la parole et du locuteur, comme peuvent le proposer Charhad *et al.* (2005); Ordelman *et al.* (2009); Vallet *et al.* (2016). La tâche de SRL est une étape de pré-traitement nécessaire pour l'identification du locuteur (Bonastre *et al.*, 2000) ou la transcription de parole (Anguera *et al.*, 2012) dans des émissions télévisées. Les tâches de SRL et d'identification permettent de déterminer « qui parle quand ». Les systèmes de SRL sont généralement fondés sur des méthodes de segmentation et regroupement non supervisées, estimant le nombre de locuteurs et découpant le flux audio en segments de parole étiquetés par des labels anonymes. Cependant, les systèmes de SRL à l'état de l'art ne sont toujours pas suffisamment précis pour être employés tels quels dans les applications de l'Ina, principalement à cause de la large variété des collections. La variété porte sur la période temporelle qui va de la fin du dix-neuvième siècle à nos jours, le type d'émissions ou les conditions d'enregistrement. Les interventions humaines sont donc la plupart du temps requises pour obtenir des annotations robustes. De plus, l'annotation de parole entièrement manuelle ne peut pas être une solution raisonnable au regard du coût important du processus. En effet, neuf heures sont requises pour effectuer l'annotation manuelle d'une heure de parole spontanée (transcription de parole et identification des locuteurs) (Bazillon *et al.*, 2008).

Dans cet article, nous proposons un simulateur pour expérimenter des méthodes de SRL assistées par l'humain afin de réduire le coût de correction d'une segmentation en locuteurs. Plus précisément, les objectifs sont de construire un automate qui simule les corrections des annotateurs et de proposer une mesure qui évalue le coût de ces corrections. Dans cet article, nous présentons dans un premier temps l'état de l'art dans le domaine d'annotation des systèmes de reconnaissance de la parole et du locuteur. Ensuite, nous décrivons le système de SRL assisté par l'humain proposé ainsi qu'une nouvelle mesure pour évaluer de tels systèmes. Dans la partie suivante, nous définissons les actions utilisées pour corriger la SRL. Avant de conclure, nous mesurons la durée de chaque action pour construire la mesure proposée et nous proposons une évaluation reposant sur un système oracle.

## 2 Travaux précédents

L'annotation humaine d'un document audio est une tâche longue et souvent fastidieuse. Elle est généralement réalisée manuellement avec des logiciels d'annotation comme *Transcriber* (Barras *et al.*, 2001) ou *ELAN* (Wittenburg *et al.*, 2006). Dans Bazillon *et al.* (2008), les auteurs ont montré que la correction des sorties d'un système de transcription automatique de la parole permet de diminuer le temps d'annotation. Une méthode d'apprentissage actif (active learning en anglais), proposée dans Budnik *et al.* (2014), utilisée en conjonction avec des systèmes de reconnaissance automatique du locuteur et du visage, réduit davantage le nombre d'interactions homme-machine. Récemment, dans Broux *et al.* (2016), nous avons proposé un système qui assiste la SRL et réduit le nombre d'interventions humaines. Dans ce travail, l'annotateur corrige seulement les erreurs de regroupement en locuteurs et la segmentation est supposée être parfaite ainsi que sans erreurs. Les deux derniers articles cités se concentrent sur la correction des erreurs de regroupement en locuteurs et négligent les erreurs de segmentation. De plus, les auteurs supposent sans le justifier que toutes les corrections ont le même coût.

# 3 Système de SRL assisté par l'humain

## 3.1 Description du système

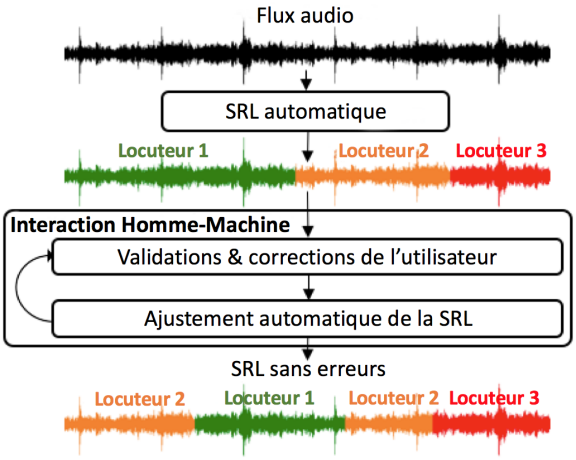



FIGURE 1 – Architecture du système de SRL assisté par l'humain

La figure 1 présente l'architecture du système de SRL assisté par un humain que nous proposons. Il est composé de deux parties principales. La première consiste à appliquer un système de SRL automatique sur un flux audio. Une segmentation initiale du flux est alors obtenue. La seconde consiste à demander à un humain de corriger la sortie de la première partie. Chaque correction humaine est à son tour prise en considération par un système qui améliore la SRL en réaffectant des segments à des locuteurs différents. Cet ajustement rend généralement plus faciles les actions restantes de l'annotateur. Selon l'objectif visé, l'annotateur peut effectuer des corrections sur le regroupement en locuteurs et/ou sur la segmentation. À la fin du processus, le taux d'erreurs de SRL, plus connu sous le nom de Diarization Error Rate (DER (NIST, 2003)), devrait avoir diminué et même être nul si toutes les erreurs de segmentation et de classification sont corrigées.

## 3.2 Simulateur expérimental

À partir de l'architecture présentée dans la précédente section, plusieurs règles ont été définies :

1. l'annotateur est simulé par un automate et ne fait aucune erreur ;
2. l'annotateur corrige l'émission du début à la fin dans l'ordre temporel afin de valider l'annotation automatique faite a posteriori ;
3. seul le tour de parole courant, qui vient d'être écouté, peut être corrigé par l'annotateur.

La première règle permet d'éviter la modélisation aléatoire et complexe des erreurs pouvant être commises par un humain. De plus, cette simplification permet d'avoir un document sans erreurs à la fin du processus de correction et ainsi un DER à 0%. Corriger à partir du début jusqu'à la fin est supposé aider la compréhension de l'annotateur et améliorer la correction. Cette règle et la dernière sont des conditions expérimentales choisies pour faciliter notre problème. Elles peuvent être remises en cause par la suite.

## 3.3 Mesure proposée : HCIQ

Le DER mesure la qualité de la segmentation et du regroupement en locuteurs (NIST, 2003). Cependant, il n'est pas pertinent pour évaluer le temps de création ou de correction d'un fichier de SRL. Une nouvelle mesure inspirée du Keystroke Saving Rate (KSR) utilisé dans la prédiction de mot pour les personnes ayant des difficultés de communication (Wood & Lewis, 1996) est proposée. Nous l'appelons Human-Computer Interaction Quantity (HCIQ). Cette mesure estime le coût des interventions d'un humain pour la correction d'une SRL. Elle peut être calculée à la fois pour les systèmes assistés et pour les systèmes où un humain corrige la SRL seul. En outre, ainsi que le DER, la mesure HCIQ peut être calculée pour chaque enregistrement ou pour un ensemble d'enregistrements audio/vidéo. Elle est définie par la formule suivante :

$$HCIQ = \sum_{i=1}^{K} w_i n_i,$$

où $i$ correspond à un type d'action de correction dans l'interface, $w_i$ est son coût associé, $n_i$ le nombre de fois que l'annotateur a appliqué ce type d'action et $K$ est le nombre de types d'action différents. Plus la mesure HCIQ est faible, plus le temps de correction de l'annotation sera faible. Au passage, elle permet de comparer différents systèmes de SRL assistés d'une manière objective pour un corpus donné.

La mesure HCIQ, en tant que telle, ne permet pas de comparer des corpus différents. Pour y remédier, nous proposons la formule suivante :

$$HCIQ_n = \frac{HCIQ}{d},$$

où $HCIQ$ est normalisé par $d$ la durée du corpus sur lequel a été calculé le $HCIQ$. Cette normalisation permet de rester dépendant de la spécifité du corpus (nombre de locuteurs, parole spontanée, etc). La mesure HCIQ$_n$ est un ratio du nombre de corrections à faire pour une unité de temps. Plus la valeur est élevée pour un corpus donné, plus ce dernier requiert des corrections.

La mesure HCIQ se rapproche de celle proposée dans Guillaumin *et al.* (2009) où les auteurs proposent une mesure reflétant l'effort nécessaire pour un humain à la correction d'images mal labelisés.

# 4 Annotateur et outils de SRL assistés

## 4.1 Logiciel d'annotation : Transcriber

Afin de choisir les actions humaines nécessaires à la correction, nous nous reposons sur *Transcriber*, un logiciel de référence dans la transcription de la parole et l'annotation. Ce logiciel permet de couper un flux audio en segments. Chaque segment correspond à une zone de parole et est associé à un nom de locuteur. Ce nom, ou label, peut être enrichi par des informations tel que le genre ou la langue native du locuteur. Dans *Transcriber*, les actions de segmentation sont "*Créer une frontière*", "*Supprimer une frontière*" et "*Déplacer une frontière*". L'action "*Créer une frontière*" ajoute une frontière en coupant un segment en deux parties, l'action "*Supprimer une frontière*" fusionne deux segments consécutifs et l'action "*Déplacer une frontière*" déplace la frontière d'un segment incorrectement positionné.

Concernant les actions de regroupement en locuteurs, *Transcriber* offre les actions "*Créer un label locuteur*" et "*Changer le label locuteur*". La première permet de créer un nouveau label locuteur pour le segment sélectionné tandis que la dernière permet de changer le label locuteur en sélectionnant un autre label parmi la liste des labels déjà créés.

### 4.2 Actions de correction

Afin de faciliter la création d'un annotateur simulé par un automate, les séries d'actions seront déterministes. Nous voulons qu'aucune action ne puisse être substituée par un ensemble d'actions fournissant la même correction. Une des actions de *Transcriber* ne remplit pas ce critère. L'action "*Déplacer une frontière*" peut être remplacée par les deux actions suivantes : "*Créer une frontière*" et "*Supprimer une frontière*".

Pour résumer, nous avons gardé deux actions pour modifier les frontières des segments et deux actions pour modifier les labels affectant le regroupement en locuteurs. En combinant ces actions, nous pouvons décrire toutes les corrections d'une manière unique. Finalement, les quatres actions sélectionnées utilisées dans la mesure HCIQ sont :

- "*Créer une frontière*",
- "*Supprimer une frontière*",
- "*Créer un label locuteur*",
- "*Changer le label locuteur*".

## 5 Expériences

### 5.1 Corpus

Les expériences ont été appliquées sur des enregistrements télévisuels de la campagne d'évaluation de 2013 du défi ANR-REPERE [1]. Les émissions télévisuelles proviennent de deux chaînes françaises (BFM et LCP). Le tableau 1 décrit le corpus utilisé dans les expériences. Le corpus est équilibré : il

| | |
|---|---|
| Nombre d'émissions | 7 |
| Nombre d'enregistrements | 28 |
| Temps d'enregistrement | 14h17 |
| Temps d'annotation | 2h57 |
| Nombre de locuteurs | 212 |

TABLE 1 – Description de REPERE test 2013

contient de la parole spontanée et préparée. Il est constitué de micros-trottoirs, de débats et d'émissions d'informations mais seule une partie des données est annotée (Kahn *et al.*, 2012).

1. http://www.defi-repere.fr/

## 5.2 Mesure de la durée des actions

Dans la section 4.2, nous avons selectionné quatre actions de correction. Maintenant, nous proposons une méthode pour estimer la durée moyenne de chaque action. L'historique des clics de la souris et des frappes du clavier dans *Transcriber* permet de déterminer indirectement les actions successives et d'évaluer précisement la durée de chaque action. Pour enregistrer cette trace d'exécution, il est nécessaire de modifier le code source de *Transcriber*. Une entrée dans le fichier des traces, c'est-à-dire un clic de souris ou une frappe de clavier, contient trois types d'information : le temps du clic ou de la frappe, le nom du module actif et un commentaire (figure 2). Le module identifie un élément de l'interface utilisateur tandis que le commentaire donne des informations précises sur l'évènement en cours. Le fichier des traces lui-même n'est pas suffisant pour déterminer les

```
[1319494751] :: [Player] [Strategy: Play/Pause; Pause at 654.942]
[1319493381] :: [LabelWindow] [Open window; Edit an existing Turn]
[1319491287] :: [Label] [Edit an existing speaker thanks to LabelWindow]
[1319480451] :: [LabelWindow] [Validate; Close Window]
```

FIGURE 2 – Exemple d'un fichier des traces

actions d'une manière automatique. En effet, l'annotateur peut faire des erreurs ou prendre une pause durant la session d'annotation contrairement à notre automate simulant un annotateur idéal. Pour résoudre ce problème, l'enregistrement de l'écran utilisateur, conservé sous la forme d'un film, est manuellement segmenté en actions en interprétant conjointement la vidéo et le fichier des traces. Chaque segment créé correspond donc précisément à la durée mesurée des actions actuelles. Afin

| **Action** | **Nombre d'occurrences** | **Moyenne (sec)** | **Écart type (sec)** |
|---|---|---|---|
| Créer un label locuteur | 28 | 12,7 | 6,0 |
| Changer le label locuteur | 32 | 7,6 | 3,8 |
| Créer une frontière | 38 | 12 | 7,6 |
| Supprimer une frontière | 46 | 5,1 | 2,3 |

TABLE 2 – Durée des actions - 20 min des données REPERE test 2013

de faciliter l'annotation de chaque action et de minimiser les erreurs d'interprétation, seules les régions avec peu de paroles spontanées et sans paroles superposées ont été annotées. Le tableau 2 montre les résultats de la durée des actions. Les actions les plus chronophages sont "*Créer un label locuteur*" et "*Créer une frontière*", avec une moyenne comprise entre 12 et 13 secondes. La première action requiert d'entrer un label locuteur (et éventuellement d'autres méta-données du locuteur), alors que la seconde action requiert de regarder et écouter le signal pour détecter la frontière de locuteurs. On notera qu'il est généralement nécessaire d'écouter le signal plusieurs fois afin de placer une nouvelle frontière. L'action nommée "*Changer le label locuteur*" a une durée moyenne de 7,6 secondes. Affichée dans une fenêtre contextuelle, elle consiste à sélectionner le label locuteur correct dans une liste déroulante. Chercher un label dans une liste déroulante demande moins d'efforts cognitifs que créer une frontière. L'action la plus rapide est l'action "*Supprimer une frontière*". Elle requiert d'arrêter l'écoute lorsqu'une frontière erronée est détectée et de la supprimer par une simple combinaison de touches au clavier.

## 5.3 Évaluation d'un système oracle

La simulation de l'annotateur repose sur deux types d'information pour déterminer si une correction est requise au temps $t$ :

1. une différence ou non entre le segment de référence (vérité terrain) et le segment de l'hypothèse ;
2. la correspondance ou non entre les labels locuteurs de la référence et l'hypothèse qui minimise le DER.

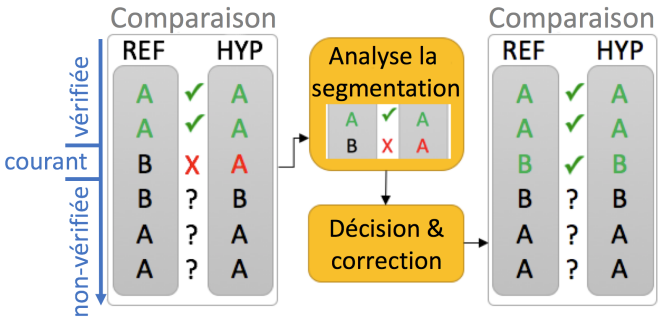



FIGURE 3 – Illustration d'un annotateur simulé

En cas de discordance au niveau de la segmentation ou du label au temps $t$ entre la référence et l'hypothèse, une correction est nécessaire. L'annotateur simulé corrige en premier les erreurs de segmentation, puis les erreurs de regroupement en locuteurs (figure 3). Après chaque correction, le système peut lancer un système de SRL sur la partie non vérifiée (les segments avec un temps de début $> t$) en prenant en considération les segments déjà vérifiés (segments avec un temps de fin $\leq t$). Dans le cas d'un système initial sans erreur de segmentation, la correction du regroupement en locuteurs est facile à mettre en place (Broux *et al.*, 2016) car les segments de l'hypothèse sont identiques aux segments de la référence. La correction de la segmentation est plus difficile, l'annotateur simulé a besoin de prendre en considération la précision des frontières de la référence. Pour résoudre ce problème, une tolérance de plus ou moins 250 ms est généralement appliquée aux frontières des segments de référence pour le calcul du DER. Nous appliquons la même tolérance pour éviter les nombreuses micro-corrections, généralement inutiles. Ainsi, avant de procéder à l'évaluation d'une possible différence entre la zone de segmentation de la référence et celle de l'hypothèse, toute frontière de l'hypothèse appartenant à une zone de tolérance est déplacée à coût nul afin d'être alignée avec la frontière de la référence.

L'annotateur simulé devient un système oracle quand aucun ajustement automatique n'est effectué au fur et à mesure des corrections. L'évaluation du système oracle est reportée dans le tableau 3. Le HCIQ du corpus test est de 331,6 minutes et correspond à la somme de toutes les estimations de durée de correction (tableau 3). La SRL utilisée comme entrée de l'oracle est fournie par le système de SRL entièrement automatique décrit dans Meignier & Merlin (2010). Le DER de la SRL avant correction est de 13,85%. Le nombre d'occurrences des actions de segmentation est environ une fois et demie plus important que le nombre d'actions de regroupement en locuteurs (respectivement 1142 et 758). Les erreurs de segmentations représentent environ 65% du temps de correction total (210,5 minutes). "*Créer une frontière*" est l'action la plus coûteuse, car elle correspond à environ 52% des corrections globales. Pour un enregistrement audio de 2h57 (177 minutes), un annotateur passera 3h17 (197,2

| Action | Nombre d'occurrences | Durée estimée (min) |
|---|---|---|
| Créer un label locuteur | 295 | 62,4 |
| Changer le label locuteur | 463 | 58,7 |
| Créer une frontière | 986 | 197,2 |
| Supprimer une frontière | 156 | 13,3 |

TABLE 3 – Correction pour le système oracle - REPERE test 2013. Durée estimée (durée moyenne $\times$ nombre d'occ.)

minutes) à créer des frontières. Si l'annotateur simulé corrige seulement les erreurs de regroupements en locuteurs, le DER est de 5,59% à la fin du processus de correction. Ces 5,59% d'erreurs sont dus à la mauvaise segmentation. Ce résultat montre que les erreurs de segmentation et de regroupement en locuteurs contribuent globalement à 40% et 60% du DER respectivement. Comparativement, les erreurs de segmentation correspondent au coût de correction principal en termes de HCIQ.

| Corpus | F | M | S | P | C | H | T | $H_n$ |
|---|---|---|---|---|---|---|---|---|
| ESTER test 2003 | 0,77 | 0,56 | 17,53 | 9,76 | 20,80 | 477,2 | 592 | 0,81 |
| ESTER test 2009 | 1,45 | 0,86 | 13,05 | 8,67 | 28,67 | 482,0 | 430 | 1,12 |
| ETAPE test 2012 | 14,93 | 0,35 | 18,74 | 9,96 | 10,16 | 793,7 | 418 | 1,90 |
| REPERE test 2013 | 0,76 | 3,60 | 9,49 | 11,47 | 8,68 | 331,6 | 177 | 1,87 |

TABLE 4 – Comparaison des $HCIQ_n$ obtenus à partir des corrections du système oracle sur divers corpus. F : $DER_{false\,alarm}(\%)$ ; M : $DER_{missed\,speaker}(\%)$ ; S : $DER_{confusion}(\%)$ ; P : Pureté (%) ; C ; Couverture (%) ; H : HCIQ (min) ; T : Temps d'annotation (min) ; $H_n$ : $HCIQ_n(min)$

Le tableau 4 permet de comparer le $HCIQ_n$ de différents corpus. Il prouve que le corpus REPERE est un des corpus qui requiert le plus de corrections pour une unité de temps car il nécessite en moyenne 1,87 minutes de corrections humaines pour 1 minute de signal audio. En outre, il montre que les corpus ETAPE et REPERE, principalement plus composés de parole spontanée (faux départs, répétitions, parole superposée, interjections, etc (Bazillon *et al.*, 2008)) que les corpus ESTER, obtiennent des scores $HCIQ_n$ élevés.

# 6 Conclusion

Dans cet article, nous avons proposé un simulateur pour évaluer des systèmes intéractifs de SRL prenant en compte les corrections humaines. La combinaison de quatre actions permet de décrire les étapes de correction d'une manière unique. Nous avons proposé une mesure pour déterminer précisément la durée de chaque action afin d'évaluer le coût des interactions homme-machine. L'évaluation du système oracle sur le corpus REPERE test 2013 montre que les corrections de segmentation prennent plus de temps que les corrections de regroupement en locuteurs. Les résultats de l'oracle montre également l'importance des erreurs de segmentation sur le HCIQ et le DER. La correction des erreurs de segmentation fait croître le HCIQ tandis qu'elle affecte de façon négligable le DER. Seule la correction des erreurs de classification fait diminuer directement le DER. Les prochains travaux se concentreront sur le développement d'un système de SRL intégré pour réduire le temps de correction.

# Références


ANGUERA X., BOZONNET S., EVANS N., FREDOUILLE C., FRIEDLAND G. & VINYALS O. (2012). Speaker diarization : A review of recent research. *ieee-tsap*, **20**(2), 356–370.

BARRAS C., GEOFFROIS E., WU Z. & LIBERMAN M. (2001). Transcriber : development and use of a tool for assisting speech corpora production. *Speech Communication*, **33**(1), 5–22.

BAZILLON T., ESTÈVE Y. & LUZZATI D. (2008). Transcription manuelle vs assistée de la parole préparé et spontanée. *Revue TAL*.

BONASTRE J.-F., DELACOURT P., FREDOUILLE C., MERLIN T. & WELLEKENS C. (2000). A speaker tracking system based on speaker turn detection for nist evaluation. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 2, p. II1177–II1180 : IEEE.

BROUX P.-A., DOUKHAN D., PETITRENAUD S., MEIGNIER S. & CARRIVE J. (2016). An active learning method for speaker identity annotation in audio recordings. In *1st International Workshop on Multimodal Media Data Analytics (MMDA), In conjuction with the 22nd European Conference on Artificial Intelligence (ECAI)*.

BUDNIK M., POIGNANT J., BESACIER L. & QUÉNOT G. (2014). Automatic propagation of manual annotations for multimodal person identification in tv shows. In *Content-Based Multimedia Indexing (CBMI), 2014 12th International Workshop on*, p. 1–4 : IEEE.

CHARHAD M., MORARU D., AYACHE S. & QUÉNOT G. (2005). Speaker identity indexing in audio-visual documents. In *Content-Based Multimedia Indexing (CBMI2005)*.

GUILLAUMIN M., VERBEEK J. & SCHMID C. (2009). Is that you ? metric learning approaches for face identification. In *Computer Vision, 2009 IEEE 12th international conference on*, p. 498–505 : IEEE.

KAHN J., GALIBERT O., QUINTARD L., CARRÉ M., GIRAUDEL A. & JOLY P. (2012). A presentation of the repere challenge. In *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on*, p. 1–6 : IEEE.

MEIGNIER S. & MERLIN T. (2010). Lium spkdiarization : an open source toolkit for diarization. In *CMU SPUD Workshop*, volume 2010.

NIST (2003). The rich transcription spring 2003 (RT-03S) evaluation plan. `http://www.itl.nist.gov/iad/mig/tests/rt/2003-spring/docs/rt03-spring-eval-plan-v4.pdf`.

ORDELMAN R., DE JONG F. & LARSON M. (2009). Enhanced multimedia content access and exploitation using semantic speech retrieval. In *Semantic Computing, 2009. ICSC'09. IEEE International Conference on*, p. 521–528 : IEEE.

VALLET F., URO J., ANDRIAMAKAOLY J., NABI H., DERVAL M. & CARRIVE J. (2016). Speech trax : A bottom to the top approach for speaker tracking and indexing in an archiving context. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)* : European Language Resources Association (ELRA).

WITTENBURG P., BRUGMAN H., RUSSEL A., KLASSMANN A. & SLOETJES H. (2006). Elan : a professional framework for multimodality research. In *Proceedings of LREC*, volume 2006, p. 5th.

WOOD M. E. & LEWIS E. (1996). Windmill-the use of a parsing algorithm to produce predictions for disabled persons. *PROCEEDINGS-INSTITUTE OF ACOUSTICS*, **18**, 315–322.

# Transcription phonétique automatique pour la synthèse de la parole

Kévin Vythelingum[1, 2] Yannick Estève[2] Olivier Rosec[1]
(1) Voxygen, Pleumeur-Bodou, France
(2) LIUM, Le Mans Université, France
{kevin.vythelingum,yannick.esteve}@univ-lemans.fr,
{kevin.vythelingum,olivier.rosec}@voxygen.fr

RÉSUMÉ

La synthèse de la parole consiste à produire un signal de parole à partir d'une séquence de mots. Elle s'appuie sur un ensemble d'enregistrements de parole transcrits en mots et en chaînes phonétiques. La qualité de cette transcription influe directement sur la qualité globale des systèmes de synthèse. Or, les chaînes phonétiques sont généralement issues d'une phonétisation automatique du texte, qui ne varie donc pas d'un locuteur à l'autre. Dans ce travail, nous explorons différentes méthodes permettant d'obtenir des chaînes phonétiques dépendantes du signal de parole et du texte. Nous appliquons finalement nos résultats à la tâche de détection des erreurs de phonétisation. Autrement dit, nous cherchons à identifier des zones où les chaînes phonétiques initiales sont erronées. Sur des données en français, nous montrons que nous pouvons corriger de 76,6 à 90,7% des erreurs de phonétisation d'un système commercial en ne vérifiant que 3,6 à 18,5% des données.

ABSTRACT

**Automatic phonemic transcription for text-to-speech synthesis**

Text-to-speech synthesis (TTS) purpose is to produce a speech signal from an input text. This implies the annotation of speech recordings with word and phonemic transcriptions. The overall quality of TTS highly depends on the accuracy of phonemic transcriptions. However, they are generally automatically produced by grapheme-to-phoneme conversion systems, which don't deal with speaker variability. In this work, we explore ways to obtain signal-dependent and context-dependent phonemic transcriptions. We then apply our results on error detection of grapheme-to-phoneme conversion hypotheses in order to find where the phonemic transcriptions may be erroneous. On a French TTS dataset, we show that we can correct from 76.6 to 90.7 % of grapheme-to-phoneme conversion errors of a commercial system by checking only 3.6 to 18.5 % of phonemes.

MOTS-CLÉS : transcription phonétique, synthèse de la parole, détection automatique d'erreurs.

KEYWORDS: grapheme-to-phoneme conversion, text-to-speech synthesis, automatic error detection.

# 1 Introduction

La synthèse de la parole consiste à produire un signal de parole à partir d'une séquence de mots. La construction d'un tel système nécessite la création d'une base de données de signal segmenté (BDS), c'est-à-dire d'un corpus composé d'un ensemble d'enregistrements de parole transcrits en mots et


10.21437/JEP.2018-12

segmentés en unités acoustiques, chacune décrite par un phonème.

Plusieurs paradigmes peuvent être distingués en synthèse de parole, dont la synthèse par corpus, où le signal de parole résulte de la sélection et de la concaténation d'unités acoustiques (Hunt & Black, 1996), et la synthèse de parole paramétrique, où les paramètres acoustiques du signal de parole sont directement prédits à partir de la description linguistique de la séquence à prononcer (Zen *et al.*, 2009). Dans le premier cas, la BDS sert de corpus de sélection des unités acoustiques. Dans le second cas, elle permet l'apprentissage des modèles acoustiques. Récemment, des travaux ont montré que certains composants des systèmes de synthèse de la parole peuvent être remplacés par des modèles neuronaux, appris grâce à des descriptions phonétiques (Van den Oord *et al.*, 2016; Arık *et al.*, 2017a,b), ou non (Wang *et al.*, 2017; Shen *et al.*, 2017).

La transcription phonétique des BDS est généralement issue d'une phonétisation automatique du texte. De nombreuses approches ont été proposées dans la littérature. Parmi celles-ci, les plus populaires sont la recherche dans un dictionnaire, la phonétisation par règles (Béchet, 2001), les modèles de séquences jointes (Bisani & Ney, 2008; Galescu & Allen, 2002) et les systèmes inspirés de la traduction automatique qui considèrent des séquences de caractères à traduire en séquences de phonèmes (Laurent *et al.*, 2009; Rao *et al.*, 2015; Yao & Zweig, 2015).

Ne dépendant que du texte, la phonétisation automatique ne varie pas selon les locuteurs et selon les situations. Lors de la création d'une BDS pour une nouvelle voix, il est donc nécessaire de procéder à des ajustements, soit au niveau du paramétrage du phonétiseur, soit en corrigeant manuellement les phonèmes qui divergent du signal constaté. Dans Brognaux *et al.* (2014), les auteurs montrent que la correction manuelle des transcriptions phonétiques de BDS en français peut améliorer la qualité de la synthèse de la parole. De plus, des améliorations de la synthèse ont été constatées dans Dall *et al.* (2016) lorsque la phonétisation est meilleure. Il est donc essentiel que la transcription phonétique des BDS soit la plus juste possible. Autrement dit, qu'elle s'accorde avec le signal de parole.

Précédemment, nous avons montré que nous pouvions détecter des erreurs de phonétisation en exploitant partiellement le signal de parole grâce à l'alignement forcé d'un lexique phonétisé par un modèle acoustique (Vythelingum *et al.*, 2017). Autrement dit, nous cherchions à identifier des zones où les chaînes phonétiques initiales étaient erronées. Dans ce travail, nous explorons différentes méthodes permettant d'obtenir des chaînes phonétiques dépendantes et indépendantes du signal de parole et du texte. Nous appliquons finalement nos résultats à la tâche de détection des erreurs de phonétisation.

L'article est organisé de la façon suivante : nous détaillons tout d'abord les différents systèmes de transcription phonétique, puis nous expliquons la tâche de détection d'erreurs et les métriques utilisées, avant de donner les résultats obtenus.

# 2 Systèmes de transcription phonétique

## 2.1 Transcription phonétique à partir du texte

### 2.1.1 Système de phonétisation par règles

Le système de phonétisation par règles (Fig. 1) est un système propriétaire utilisé pour la synthèse de la parole en français. Il prend en entrée une séquence de mots et donne en sortie la transcription

phonétique de ceux-ci.

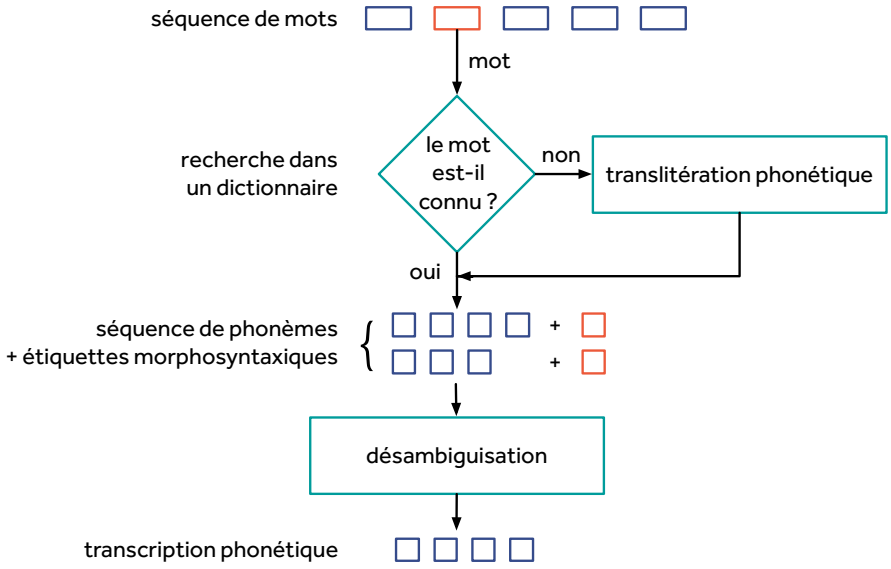



FIGURE 1 – Système de phonétisation par règles

Chaque mot est recherché individuellement dans un dictionnaire. S'il est présent, nous obtenons une ou plusieurs variantes phonétiques, associées à une étiquette morphosyntaxique. Sinon, des règles de translitération sont appliquées. Finalement, un module de désambiguisation permet de choisir une des variantes phonétiques proposées selon le contexte lexical.

### 2.1.2 Système de phonétisation par modèles de séquences jointes

Un modèle de séquences jointes est un modèle statistique permettant d'associer des séquences de lettres à des séquences de phonèmes (Bisani & Ney, 2008; Galescu & Allen, 2002). Le principe est d'effectuer un alignement des séquences de lettres, appelés graphèmes, et des phonèmes. Les séquences de couples *(graphèmes, phonèmes)* sont ensuite modélisées par un modèle de langage. La phonétisation d'un mot se fait donc en deux étapes : la génération des différentes séquences de couples *(graphèmes, phonèmes)* possibles selon les graphèmes du mot, puis la désambiguisation des variantes phonétiques grâce au modèle de langage.

Nous utilisons l'outil Phonétisaurus (Novak *et al.*, 2012, 2013) pour le modèle d'alignement phonétique. Ce dernier est associé à un modèle de language 6-gramme construit avec SRILM (Stolcke, 2002; Stolcke *et al.*, 2011) sur l'alignement graphèmes-phonèmes du corpus d'apprentissage.

Contrairement au système de phonétisation par règles, le système de phonétisation par modèles de séquences jointes ne peut traiter que des mots isolés. Il nous est donc davantage utile pour enrichir un lexique phonétisé que pour directement annoter en phonèmes les BDS.

### 2.1.3 Système de phonétisation par réseau de neurones

Pour obtenir une hypothèse de transcription phonétique dépendante du contexte lexical des mots, nous avons développé un modèle neuronal fondé sur l'architecture encodeur-décodeur décrite dans

(Bahdanau *et al.*, 2015). L'encodeur est un réseau de neurones récurrent avec une couche de GRU (Gated Reccurent Unit) de taille 128 bidirectionnelle. Il prend en entrée des mots segmentés au niveau caractères projetés dans un espace à 64 dimensions. Le décodeur est quant à lui composé de deux couches de GRU avec un mécanisme d'attention.

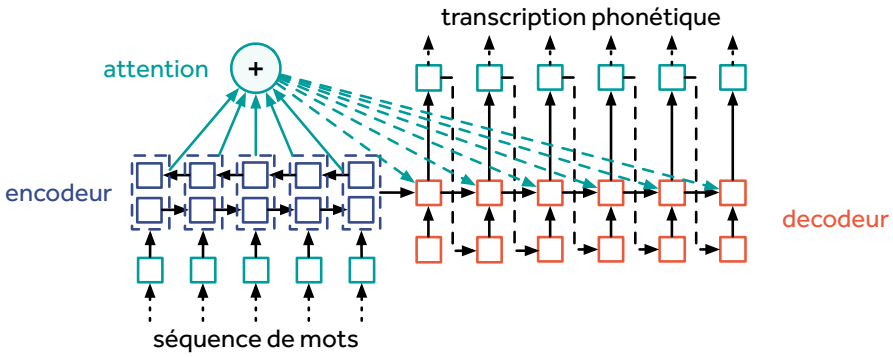



FIGURE 2 – Système de phonétisation par réseau de neurones

Nous utilisons l'outil Nmtpy (Caglayan *et al.*, 2017) dans sa configuration par défaut pour mettre en oeuvre ce modèle de traduction automatique neuronale. Ceci nous permet d'obtenir une hypothèse de phonétisation dépendante du texte sans réaliser d'alignement préalable entre caractères et phonèmes.

## 2.2 Transcription phonétique par alignement forcé

Pour prendre en compte le signal acoustique dans l'annotation en phonèmes des BDS, nous procédons à un alignement forcé du texte sur la parole. Ceci est fait grâce à un modèle acoustique d'une part, et à un lexique phonétisé d'autre part.

Tout d'abord, nous entraînons un modèle acoustique HMM-GMM (Hidden Markov Models - Gaussian Mixture Models) sur des paramètres PLP (Perceptual Linear Prediction) avec une adaptation au locuteur fMLLR (feature space Maximum Likelihood Linear Regression). Ensuite, nous entraînons un modèle HMM-DNN (Hidden Markov Models - Deep Neural Networks) sur les mêmes paramètres acoustiques en utilisant l'alignement produit par le modèle HMM-GMM. Notre modèle acoustique neuronal est composé d'une couche d'entrée de 360 dimensions, correspondant à des paramètres acoustiques de 40 dimensions concaténés aux paramètres voisins sur une fenêtre de 8 trames). De plus, 5 couches cachées comportant 3000 dimensions et une couche de sortie de 10553 dimensions composent le modèle. Les modèles acoustiques sont construits grâce à Kaldi (Povey *et al.*, 2011) sur des BDS existantes.

Le lexique utilisé pour l'alignement forcé est issu de la transcription phonétique des mots du corpus de test par le système de phonétisation par règles. Comme certaines hypothèses de prononciation sont manquantes, nous enrichissons ce lexique avec des hypothèses du système de phonétisation par modèles de séquences jointes. Afin de déterminer le nombre optimal d'hypothèses de prononciation à ajouter par mot, nous réalisons plusieurs alignements forcés avec à chaque fois un lexique plus ou moins enrichi.

## 2.3 Transcription phonétique à partir du signal de parole

### 2.3.1 Système de reconnaissance de phonèmes HMM-DNN

L'alignement forcé dépendant des variantes de prononciations présentes dans le lexique, le modèle acoustique est limité pour choisir l'hypothèse de phonétisation la plus probable. Une autre manière de prendre en compte le signal de parole dans la transcription phonétique des BDS est de réaliser une reconnaissance de phonèmes sur la partie acoustique des données.

Le système de reconnaissance de phonèmes utilise le même modèle acoustique que pour l'alignement forcé, le lexique étant constitué de la liste des phonèmes. Pour le décodage, nous avons appris un modèle de langage trigramme au niveau phonèmes sur le corpus d'apprentissage du modèle acoustique. Ce dernier ayant été validé manuellement, il nous permet d'apprendre les contraintes phonotactiques sur des données fiables.

### 2.3.2 Système de reconnaissance de phonèmes de bout en bout

Un système de reconnaissance de phonèmes de bout en bout est également évalué. Il s'agit de l'architecture de Deepspeech 2 (Amodei *et al.*, 2015), composée d'un réseau de neurones avec 2 couches de convolution et 5 couches de GRU bidirectionnelles de taille 800. Le système est entraîné grâce à la fonction d'activation CTC (Connectionist Temporal Classification) (Graves *et al.*, 2006) afin d'éviter la phase d'alignement entre les séquences de phonèmes et le signal. En effet, comme l'alignement temporel n'est pas requis pour la tâche de transcription phonétique, nous pouvons alors éviter d'induire des erreurs inhérentes à cette tâche.

# 3 Détection des erreurs de phonétisation

Nous cherchons à détecter les erreurs de phonétisation du système à base de règles. Pour cela, nous comparons ses sorties à des transcriptions corrigées manuellement, afin d'annoter l'ensemble des phonèmes avec des étiquettes *correct* ou *erreur*. Nous obtenons donc la référence pour la détection d'erreurs. Ensuite, les hypothèses des différents systèmes de transcription phonétique sont alignées avec les sorties du phonétiseur à base de règles : des phonèmes différents donneront une étiquette *erreur*, tandis que des phonèmes identiques donneront l'étiquette *correct*. Nous obtenons ainsi les différentes hypothèses de détection d'erreurs. Finalement, nous comparons référence et hypothèses de détections d'erreurs pour évaluer cette tâche.

Plusieurs métriques sont utilisées pour l'évaluation. D'une part, nous utilisons *précision* et *rappel* pour déterminer respectivement la proportion de vraies alarmes et la proportion d'erreurs détectées. D'autre part, nous calculons le pourcentage de données à valider, c'est-à-dire la proportion de phonèmes qu'un annotateur doit vérifier manuellement pour corriger l'ensemble des erreurs détectées. Ce dernier correspond au rapport entre le nombre de phonèmes supposés erronés et le nombre de phonèmes de la référence :

$$\text{données à valider} = \frac{\text{nombre de phonèmes supposés erronés}}{\text{nombre de phonèmes de la référence}}$$

Nous cherchons à maximiser précision et rappel, tandis que nous cherchons à minimiser la quantité de données à valider.

# 4 Résultats

Les données que nous avons utilisées pour l'évaluation de nos systèmes sont issues des bases de données de synthèse de Voxygen. Les modèles sont appris sur environ 50 heures de parole énoncées par 9 locuteurs, soit 90 135 séquences de mots. Les modèles sont testés sur environ 10 heures de parole énoncées par 3 locuteurs, soit 16 328 séquences de mots. Les locuteurs du corpus de test ne sont pas présents dans le corpus d'apprentissage. L'un d'eux, noté *locuteur 1* dans la suite, a la particularité d'avoir un accent africain sénégalais Bien que son accent diffère des autres locuteurs au niveau acoustique, il n'induit pas l'application de règles de phonétisation différentes. Ainsi, cela nous permettra d'évaluer la robustesse de nos modèles acoustiques.

## 4.1 Évaluation des systèmes de transcription phonétique

Tout d'abord, nous évaluons la performance des systèmes de transcription phonétique sur la tâche de phonétisation. Celle-ci est donnée en terme de taux d'erreur de phonétisation, pour chaque locuteur et pour l'ensemble des données (Tab. 1). Il s'agit de comptabiliser les substitutions, insertions et omissions.

| # | Système | Locuteur 1 | Locuteur 2 | Locuteur 3 | Total |
|---|---|---|---|---|---|
| (0) | phonétisation par règles | 0,8 | 2,3 | 1,3 | 1,8 |
| (1) | phonétisation par modèles de séquences jointes | 8,4 | 9,0 | 7,4 | 8,5 |
| (2) | phonétisation par réseau de neurones | **0,1** | **2,1** | **0,5** | **1,4** |
| (3) | alignement forcé lexique de base | 5,0 | 2,9 | 2,7 | 3,1 |
| (4) | alignement forcé lexique de base + 1 variante | **4,9** | **2,4** | **1,8** | **2,5** |
| (5) | alignement forcé lexique de base + 2 variantes | 7,5 | 3,0 | 2,2 | 3,3 |
| (6) | alignement forcé lexique de base + 3 variantes | 10,1 | 3,5 | 3,0 | 4,2 |
| (7) | reconnaissance de phonèmes HMM-DNN | 54,6 | 13,3 | 12,4 | 18,3 |
| (8) | reconnaissance de phonèmes CNN-RNN | **49,3** | **11,9** | **10,9** | **16,4** |

TABLE 1 – Taux d'erreur de transcription phonétique (%) des différents systèmes étudiés

La transcription phonétique à partir du texte donne les meilleurs résultats, excepté pour la phonétisation par modèles de séquences jointes, qui ne prend pas en compte le contexte lexical des mots. La phonétisation par réseau de neurones améliore même la transcription phonétique par rapport à celle obtenue à base de règles. Les erreurs corrigées lors de l'étape de validation manuelle des données composant le corpus d'apprentissage du modèle ont donc été capturées.

Pour l'alignement forcé, nous faisons varier de 0 à 3 le nombre de variantes de prononciation que nous ajoutons à un lexique de base. Le lexique de base est constitué avec le phonétiseur par règles tandis que les variantes additionnelles sont produites par le phonétiseur par modèles de séquences jointes. Nous observons que la meilleure transcription est obtenue en ajoutant une seule variante. Cela montre que certaines variantes utiles n'ont pas été produites par le phonétiseur par règles mais qu'une trop grande variabilité dans le lexique nuit à l'alignement forcé.

Concernant la reconnaissance de phonèmes, le modèle neuronal de bout en bout permet de gagner deux points de taux d'erreur sur le modèle HMM-DNN, et ce pour chaque locuteur. Bien que le taux d'erreur soit beaucoup plus élevé que pour les systèmes dépendant du texte, il est possible que les erreurs soient différentes des autres systèmes, et permettent de détecter certaines erreurs de phonétisation.

Nous observons finalement que le taux d'erreur est plus élevé pour le locuteur 1 pour les systèmes dépendants du signal de parole. Les modèles acoustiques utilisés ne sont donc pas robustes à la particularité de l'accent de ce locuteur.

## 4.2 Application à la détection des erreurs de phonétisation

Nous évaluons les systèmes de transcription phonétique sur la tâche de détection des erreurs de phonétisation du système à base de règles. Les résultats sont donnés en termes de précision, rappel, et de quantité de données à valider pour corriger les erreurs détectées. Ceci permet d'estimer quel effort la validation nécessitera et quelle quantité d'erreurs pourra-t-on espérer corriger.

Dans un premier temps, nous considérons les différents systèmes de manière isolée (Tab. 2).

| # | Système | Précision | Rappel | Données à valider |
|---|---|---|---|---|
| (1) | phonétisation par modèles de séquences jointes | 13,4 | **64,7** | 8,6 |
| (2) | phonétisation par réseau de neurones | **68,4** | 58,5 | **1,5** |
| (3) | alignement forcé lexique de base | 30,9 | 51,0 | 3,0 |
| (4) | alignement forcé lexique de base + 1 variante | **39,8** | 64,8 | **2,9** |
| (5) | alignement forcé lexique de base + 2 variantes | 32,5 | 72,7 | 4,0 |
| (6) | alignement forcé lexique de base + 3 variantes | 27,0 | **75,4** | 5,0 |
| (7) | reconnaissance de phonèmes HMM-DNN | 7,8 | **85,2** | 19,3 |
| (8) | reconnaissance de phonèmes CNN-RNN | **8,5** | 83,0 | **17,3** |

TABLE 2 – Évaluation des systèmes étudiés sur la tâche de détection des erreurs de phonétisation (%)

Nous observons que les systèmes permettant la meilleure précision et offrant la plus petite quantité de données à valider sont ceux qui obtenaient les taux d'erreurs de phonétisation les plus faibles. Cependant, le meilleur rappel est obtenu à l'inverse par ceux dont la transcription divergeaient le plus de la référence. Nous comprenons qu'il est nécessaire de trouver un compromis entre temps passé et nombre d'erreurs corrigées lors de la validation manuelle. Les différents systèmes permettent donc de cibler différents niveaux de qualité de transcription. Nous remarquons ainsi que valider 1,5% des données permet de corriger 58,5% des erreurs, valider 2,9% des données permet de corriger 64,8% des erreurs, valider 5,0% des données permet de corriger 75,4% des erreurs et valider 17,3% des données permet de corriger 83,0% des erreurs.

Dans un second temps, nous considérons la combinaison des hypothèses de phonétisation des meilleurs systèmes étudiés (Tab. 3). Nous avons trois systèmes, soit celui qui obtenait le plus faible taux d'erreur de transcription phonétique pour chaque source utilisée. Nous profitons ainsi de la complémentarité des transcriptions phonétiques issues du signal de parole et du texte.

Nous observons que la combinaison de l'alignement forcé avec la phonétisation à partir du texte permet un gain important en rappel pour la même précision par rapport à l'alignement forcé seul. Nous passons en effet de 64,8% à 76,6% de rappel. De plus, la combinaison de la reconnaissance

| Systèmes combinés | Précision | Rappel | Données à valider |
|---|---|---|---|
| (2) + (4) | **38,4** | 76,6 | **3,6** |
| (2) + (8) | 8,8 | 87,2 | 17,5 |
| (4) + (8) | 8,5 | 87,7 | 18,1 |
| (2) + (4) + (8) | 8,6 | **90,7** | 18,5 |

TABLE 3 – Évaluation de la combinaison de systèmes sur la tâche de détection des erreurs de phonétisation (%)

de phonèmes avec les autres systèmes permet les rappels les plus importants. En combinant les trois systèmes retenus, nous pouvons corriger 90,7% des erreurs en validant 18,5% des données.

# 5 Conclusion

Nous comparons plusieurs méthodes pour obtenir une transcription phonétique à partir de signal de parole et de texte. Les systèmes étudiés, bien qu'ayant des taux d'erreurs de transcription différents, permettent d'obtenir des hypothèses complémentaires. Nous proposons d'appliquer ces résultats dans le cadre de la détection des erreurs de phonétisation, une tâche très utile en synthèse de parole pour accélérer le processus de développement de nouvelles voix. En combinant les hypothèses de phonétisation complémentaires des différents systèmes étudiés, nous parvenons à identifier des zones dans les données annotées où la transcription phonétique semble erronée. Dans le but d'augmenter la qualité des bases de données de synthèse, un annotateur humain peut donc en un temps minimum améliorer globalement les transcriptions. En effet, sur des données en français, nous montrons que nous pouvons corriger de 76,6 à 90,7% des erreurs de phonétisation d'un système commercial en ne vérifiant que 3,6 à 18,5% des données. Dans un prochain travail, nous chercherons à valider notre approche sur d'autres langues, notamment celles disposant de moins de ressources en termes de données et de connaissances linguistiques.

# Références


AMODEI D., ANUBHAI R., BATTENBERG E. *et al.* (2015). Deep speech 2 : End-to-end speech recognition in english and mandarin. *CoRR*, **abs/1512.02595**.

ARIK S., CHRZANOWSKI M., COATES A. *et al.* (2017a). Deep voice : Real-time neural text-to-speech. In *arXiv :1702.07825v2*.

ARIK S., DIAMOS G., GIBIANSKY A. *et al.* (2017b). Deep voice 2 : Multi-speaker neural text-to-speech. In *arXiv :1705.08947v1*.

BAHDANAU D., CHO K. & BENGIO Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

BÉCHET F. (2001). Lia phon : un système complet de phonétisation de textes. In *Traitement Automatique des Langues (TAL)*, p. 47–67.

BISANI M. & NEY H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. In *Speech Communication*, p. 434–451.

BROGNAUX S., B. P., DRUGMAN T. & D. L. (2014). Speech synthesis in various communicative situations : Impact of pronunciation variations. In *Proceedings of InterSpeech*.

CAGLAYAN O., GARCÍA-MARTÍNEZ M., BARDET A. *et al.* (2017). Nmtpy : A flexible toolkit for advanced neural machine translation systems. *Prague Bull. Math. Linguistics*, **109**, 15–28.

DALL R., BROGNAUX S., RICHMOND K. *et al.* (2016). Testing the consistency assumption : Pronunciation variant forced alignment in read and spontaneous speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5155–5159.

GALESCU L. & ALLEN J. F. (2002). Pronunciation of proper names with a joint n-gram model for bi-directional grapheme-to-phoneme conversion. In *Proceedings of InterSpeech*.

GRAVES A., FERNÁNDEZ S., GOMEZ F. & SCHMIDHUBER J. (2006). Connectionist temporal classification : labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine learning*, p. 369–376.

HUNT A. J. & BLACK A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, p. 373–376.

LAURENT A., DELÉGLISE P. & MEIGNIER S. (2009). Grapheme to phoneme conversion using an smt system. In *Proceedings of InterSpeech*.

NOVAK J. R., DIXON P. R., MINEMATSU N. *et al.* (2012). Improving wfst-based g2p conversion with alignment constraints and rnnlm n-best rescoring. In *Proceedings of InterSpeech*.

NOVAK J. R., MINEMATU N. & HIROSE K. (2013). Failure transitions for joint n-gram models and g2p conversion. In *Proceedings of InterSpeech*.

POVEY D., GHOSHAL A., BOULIANNE G. *et al.* (2011). The kaldi speech recognition toolkit. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

RAO K., PENG F., SAK H. & BEAUFAYS F. (2015). Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 4225–4229.

SHEN J., PANG R., WEISS R. J. *et al.* (2017). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *arXiv :1712.05884*.

STOLCKE A. (2002). Srilm – an extensible language modeling toolkit. In *Proceedings of InterSpeech*.

STOLCKE A., ZHENG J. & WANG W. (2011). Srilm at sixteen : Update and outlook. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*.

VAN DEN OORD A., DIELEMAN S., ZEN H. *et al.* (2016). Wavenet : A generative model for raw audio. In *arXiv :1609.03499v2*.

VYTHELINGUM K., ESTÈVE Y. & ROSEC O. (2017). Error detection of grapheme-to-phoneme conversion in text-to-speech synthesis using speech signal and lexical context. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop*.

WANG Y., SKERRY-RYAN R., STANTON D. *et al.* (2017). Tacotron : Towards end-to-end speech synthesis. In *arXiv :1703.10135*.

YAO K. & ZWEIG G. (2015). Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. In *Proceedings of InterSpeech*.

ZEN H., TOKUDA K. & BLACK A. W. (2009). Statistical parametric speech synthesis. In *Speech Communication*, p. 1039–1064.

# Analyse électromyographique de la production des plosives labiales : enjeux méthodologiques.

Thibault Cattelain, Maëva Garnier, Christophe Savariaux, Silvain Gerber, Pascal Perrier
Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-lab, 38000 Grenoble, France
* Institute of Engineering Univ. Grenoble Alpes
thibault.cattelain@gipsa-lab.grenoble-inp.fr

## RESUME

Notre objectif est de montrer l'intérêt d'une d'analyse temporelle détaillée des gestes de production des consonnes occlusives labiales, en distinguant différentes phases, y inclus durant l'occlusion où les lèvres sont immobiles. Quatre adultes ont été enregistrés pour cette étude. Il leur était demandé de produire des logatomes contenant les plosives /p/ et /b/, dans différents contextes vocaliques (/a/, /i/, /u/) et avec différents niveaux d'effort articulatoire. Les signaux EMG de surface de cinq muscles péri-oraux ont été acquis simultanément avec une vidéo des lèvres. Les mouvements de parole ont été segmentés en quatre phases à l'aide du signal cinématique de distance interlabiale. Des patrons d'activation musculaire semblables ont été observés chez les quatre locuteurs, quels que soient la consonne et le contexte vocalique. Enfin, les activités du muscle dépresseur de la lèvre inférieure et du mentalis dans certaines phases du mouvement semblent être de bons descripteurs de l'effort articulatoire.

## ABSTRACT

Our goal is to show the interest of distinguishing several phases in the gesture underlying the production of labial consonants, including during the occlusion, where lips are essentially static, in order to better understand their control. Four adult speakers were recorded, while producing logatoms including the consonants /p/ and /b/, followed by 3 vowels (/a/, /i/ or /u/) and with increasing levels of articulatory effort. Five surface EMG signals targeting five orofacial muscles involved in several lip movements were recorded simultaneously with a video of the lips. The speech gestures were segmented into four phases, based on the interlabial distance. Similar muscle activation patterns were observed for the four speakers, regardless of the consonant (/p/ or /b/) and the vowel context. Finally, the activities of the lower lip depressor and the mentalis muscles in some movement phases appear to be reliable descriptors of articulatory effort.

MOTS-CLES : plosives, électromyographie, effort articulatoire, patrons d'activation musculaire
KEYWORDS: labial stop consonant, electromyography, articulatory effort, muscle activation patterns

# 1 Introduction

La production des gestes de parole requiert une coordination complexe des gestes respiratoires, laryngés et articulatoires. Le déplacement des articulateurs (langue, lèvres…), en particulier, est contrôlé à partir du recrutement précis de plusieurs muscles orofaciaux. La connaissance détaillée


10.21437/JEP.2018-13

des muscles impliqués dans ces gestes, et des efforts physiologiques associés, serait d'un grand intérêt pour la modélisation de la production et la compréhension de divers troubles articulatoires. Cependant, si diverses méthodes sont communément utilisées en phonétique expérimentale pour caractériser le déplacement cinématique des articulateurs, l'électromyographie de surface reste encore très rarement utilisée pour caractériser les activités musculaires orofaciales sous-jacentes à ces gestes articulatoires.

Une des principales raisons à cela est certainement la difficulté de positionnement des électrodes. L'anatomie du visage est en effet fine et complexe: la densité musculaire est élevée, avec plusieurs muscles se superposant en particulier sur le pourtour des lèvres, et l'anatomie présente une importante variabilité inter-individuelle. De ce fait, les signaux EMG enregistrés avec les techniques d'électromyographie de surface bipolaire peuvent varier considérablement en fonction du positionnement des électrodes (voir par exemple : Beck et al., 2008 ; Campanini et al., 2007 ; Hogrel et al., 1998; Jensen et al., 1993; Roy et al., 1986), et une électrode peut capter l'activité d'un muscle proche, même s'il n'est pas directement situé sous cette électrode (on parle alors de « diaphonie »). O'dwyer et al. (1981) ont été les premiers à proposer des recommandations précises pour le placement des électrodes EMG sur le visage. Ils ont pu localiser précisément l'emplacement de fibres musculaires sous le derme du visage en réalisant l'acquisition des signaux EMG avec une technique intra-musculaire très invasive, mais permettant de capter les potentiels d'action moteurs directement sur les fibres musculaires, et en demandant à leurs participants de réaliser un ensemble de mouvements orofaciaux simples, silencieux, impliquant des recrutements musculaires bien identifiés. Leurs travaux se sont concentrés sur 10 muscles orofaciaux, parmi lesquels l'orbicularis oris, le mentalis (ou mentonnier) et le dépresseur de la lèvre inférieure, principalement recrutés dans les gestes de parole. Plus récemment l'avènement de l'électromyographie à haute-densité (HD-EMG), qui exploite des matrices d'électrodes EMG de surface, a permis à Lapatki et al. (2010) d'apporter des compléments précieux aux connaissances sur la localisation des muscles orofaciaux et leur recrutement dans divers mouvements. Malheureusement, la forte variabilité anatomique inter-individuelle ne permet pas d'établir des règles universelles pour le placement des électrodes sur le visage, et des ajustements sont nécessaires pour chaque sujet lors du recueil des données.

Une autre source de difficultés relative à l'étude des activités musculaires orofaciales est la complexité de ces gestes et le fait qu'ils se composent en réalité de plusieurs sous-mouvements. La question se pose, par conséquent, de l'échelle de temps sur laquelle analyser les signaux EMG, et de la globalité vs. localité des descripteurs à en extraire. Certains auteurs (McClean et Tasko, 2003. Blair et Smith, 1986) se sont penchés sur cette question et ont ainsi cherché à décrire les patrons temporels d'activation des muscles orofaciaux lors de la production de la parole ou de mimiques faciales. McClean et Tasko (2003) ont ainsi proposé de décomposer les signaux de déplacement de la mandibule et des lèvres en une séquence de mouvements, grâce d'abord à la détection des pics de vitesse tangentielle, puis par le repérage des débuts et fins de mouvements correspondant aux minima de la vitesse tangentielle situés immédiatement avant et immédiatement après chacun des pics de vitesse. Plutôt que de mesurer l'énergie du signal EMG sur l'intégralité du mouvement, ils ont alors proposé de mesurer l'intensité électromyographique à partir du pic dans l'enveloppe EMG précédant un pic de vitesse de la mandibule ou des lèvres.

Dans cette présente étude des activations musculaires associées à la production des plosives labiales, nous proposons une méthodologie de segmentation des signaux cinématiques plus fine encore, dans laquelle nous définissons des phases correspondant à des portions de mouvements. Nous appliquons cette technique d'analyse pour décrire les patrons d'activation musculaire orofaciale (5 muscles) lors de la production de plosives prononcées par 4 sujets dans plusieurs conditions expérimentales. Dans une première partie nous présenterons le corpus et la méthode d'acquisition puis de traitement des données. Puis nous illustrerons l'intérêt de la segmentation en

phases que nous nous proposons, d'abord par une analyse globale, puis en étudiant l'effet du contexte vocalique et de l'effort articulatoire sur les activations musculaires. Nous discuterons finalement ces résultats et conclurons.

# 2 Matériel et Méthodes

## 2.1 Sujets et corpus

Quatre sujets adultes (2 femmes, âge entre 21 et 33 ans) ont participé à l'expérience. Il leur était demandé de produire six logatomes : /ləpa/, /ləba/, /ləpi/, /ləbi/, /ləpu/, /ləbu/. Chaque logatome a été répété de manière isolée par séries de 5 items d'effort articulatoire (auto-évalué) croissant, à un rythme confortable et en insérant une courte pause entre chaque item. Les enregistrements duraient 30 secondes, permettant la production de cinq à six séries de logatomes à chaque fois.

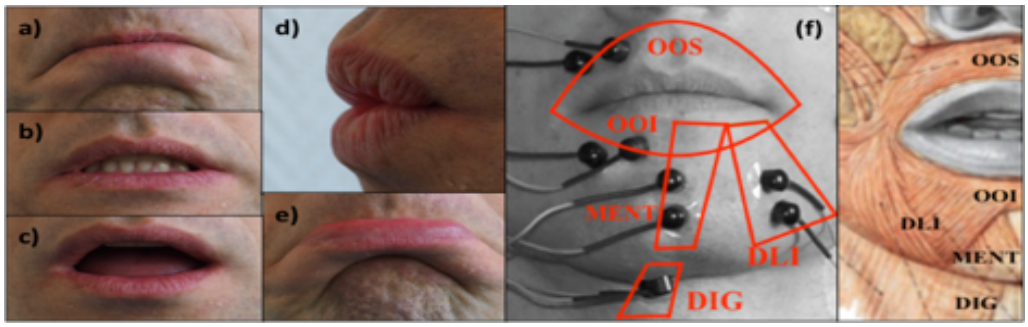



FIGURE 1 : Gestes orofaciaux silencieux utilisés pour valider le positionnement des électrodes EMG (a. compression labiale ; b.abaissement de la lèvre inférieure ; c. abaissement mandibulaire ; d. protrusion labiale ; e. relèvement de la lèvre inférieure), localisation des 5 paires d'électrodes (f) et vue anatomique dans la région des lèvres.

## 2.2 Matériel d'acquisition

Cinq signaux EMGs de surface (fe = 20kHz) ont été simultanément acquis ciblant cinq muscles du pourtour des lèvres supposés intervenir dans le geste de production des consonnes occlusives labiales (Digastrique - DIG, Mentalis - MENT, Dépresseur de la lèvre inférieure – DLI, Orbicularis oris superior et inferior – OOS et OOI, cf Figure 1.f). Les conséquences du recrutement de ces muscles ont été abondamment décrites dans la littérature (voir par exemple O'dwyers et al., 1981). Le Mentalis (MENT) est recruté pour le relèvement de la lèvre inférieure tandis que le Dépresseur (DLI) est recruté pour l'abaissement de la lèvre inférieure. Ces deux muscles agissent de manières antagonistes lors du déplacement de la lèvre inférieure. Les deux parties, supérieure et inférieure, du muscle Orbicularis Oris (OOS et OOI) sont recrutées dans les mouvements de protrusion et de compression labiale. Le Digastrique (DIG), lui, est recruté dans l'abaissement de la mandibule. Les lèvres des sujets, maquillées en bleu pour faciliter la détection de leurs contours (Lallouache, 1991), ont été filmées de face à l'aide d'une caméra rapide (100 images/s). Le signal acoustique a été acquis à l'aide d'un microphone et d'un amplificateur de mesure Bruël & Kjaer (Fréquence d'échantillonnage à 20kHz) permettant de réaliser une mesure calibrée de l'intensité acoustique. Le signal audio et les signaux EMGs ont été acquis à l'aide du système BIOPAC MP160, qui par ailleurs assure la synchronisation de tous les signaux par l'envoi d'une impulsion de synchronisation à l'ensemble des dispositifs d'acquisition.

## 2.3 Segmentation des signaux cinématiques en 4 phases et extraction des descripteurs cinématiques.

Dans la lignée des travaux de McClean & Tasko (2003) nous avons procédé à une segmentation sur la base d'évènements mesurables sur les signaux cinématiques. Mais alors que McClean & Tasko se sont essentiellement appuyés sur l'extraction des minima et maxima de la vitesse tangentielle, nous avons basé cette segmentation sur une analyse de la distance interlabiale et la détection des phases de contact entre les lèvres. En effet, lors de la production de consonnes occlusives labiales, il existe toute une phase pendant l'occlusion du conduit vocal où les lèvres sont en contact et où la vitesse de déplacement des lèvres est nulle ou très faible, alors même que l'activité musculaire est significative. Ce contact interrompt brutalement le mouvement, alors même que, de manière sous-jacente, l'activation des muscles varie pour assurer l'occlusion et pour préparer le geste d'ouverture. C'est pourquoi nous avons observé la compression des lèvres durant la phase de contact et repéré des évènements temporels sur la base de la variation de la distance interlabiale et non du mouvement des lèvres. Nous avons préféré cette démarche à celle qui aurait consisté à segmenter directement les signaux EMG car elle fournit une information sur l'organisation temporelle des gestes phonétiquement pertinents, et nous semble donc bien caractériser les buts moteurs de la tâche pour lesquels les activations musculaires sont coordonnées.

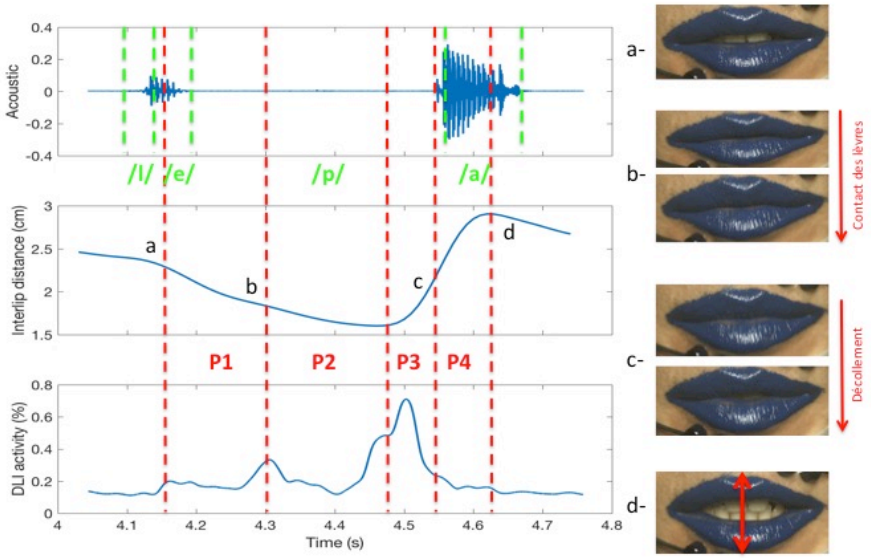



FIGURE 2 : Décomposition en 4 phases du mouvement labial lors de la production de la consonne /p/ dans /ləpa/. Les lignes pointillées vertes indiquent les limites de l'étiquetage phonétique de la séquence à partir du signal acoustique (cadre du haut); les lignes pointillées rouges marquent les évènements cruciaux pour la délimitation des phases, repérés sur le signal de distance interlabiale (cadre du mlieu): a=début de décroissance avant l'occlusion, b=début de contact, c=fin de contact, d fin de croissance après l'occlusion. Le cadre du bas représente l'activité EMG du Dépresseur de la Lèvre Inférieure (DLI).

Pour cela, dans chaque image vidéo, le contour externe des lèvres a été extrait à l'aide d'une technique d'apprentissage statistique de la couleur qui reprend les principes proposés par Lallouache (1991). Ces contours nous ont permis de mesurer la distance interlabiale qui sépare l'arc de cupidon du point le plus bas de la lèvre inférieure (Fig 2, flèche rouge sur l'image « d » et tracé du milieu). Nous avons ensuite repéré automatiquement l'instant de début de décroissance de la distance interlabiale précédant l'occlusion consonantique (Fig. 2, étiquette et image « a ») et l'instant de fin de croissance suivant le relâchement de la consonne (Fig. 2, étiquette « d »). Puis dans l'intervalle [a d] nous avons automatiquement détecté l'instant où la distance interlabiale atteint son minimum, et manuellement repéré les instants de début (Fig.2, étiquette et images « b ») et de fin (Fig2.

Etiquette et images « c ») de la phase de contact interlabial. Ces quatre étiquettes et le minimum de la distance interlabiale définissent 4 phases dans le mouvement des lèvres qui est à la base de la production de la consonne : une phase P1 de fermeture vers l'occlusion (intervalle [a b]), puis, pendant l'occlusion, une phase P2 de compression (de l'étiquette « b » au minimum de la distance interlabiale) et une phase P3 de relâchement de la compression (du minimum de la distance interlabiale à l'étiquette « c »), et enfin, après l'occlusion, une phase P4 d'ouverture (intervalle [c d]).

Trois autres paramètres ont été mesurés sur la variation de la distance interlabiale : la valeur minimum de la distance, appelée « degré de compression » et les amplitudes des pics de vitesse situés de part et d'autre de la phase d'occlusion (vitesses de fermeture et de réouverture).

## 2.4 Post-traitement des signaux EMG et mesure de l'activité musculaire par phase du mouvement

Le spectre fréquentiel des signaux EMGs est basse fréquence (maximum 500Hz). C'est pourquoi ces signaux, acquis à la fréquence d'échantillonnage de 20 kHz, ont été sous-échantillonnés à 2 kHz. Ils ont ensuite été filtrés à l'aide d'un filtre passe-haut (fc = 20Hz), afin de réduire les composantes basse-fréquences dues aux artefacts de mouvements des électrodes sur la peau. Leur enveloppe redressée a ensuite été calculée sur 500 points.

Pour chaque enregistrement (comprenant donc 20-25 répétitions d'un même logatome avec des variations d'effort), l'amplitude moyenne du bruit de mesure a été estimée sur chaque signal EMG à partir des premières centaines de millisecondes où le sujet était au repos, puis elle a été soustraite de l'enveloppe redressée du signal. Dans un second temps, pour chacun des cinq muscles ciblés, nous avons calculé la distribution des amplitudes de l'enveloppe redressée du signal EMG sur l'ensemble du corpus, repéré le seuil en dessous duquel les activités enregistrées pouvaient s'apparenter à un résidu de bruit, et soustrait cette valeur à l'enveloppe redressée du signal EMG. Les enveloppes redressées et seuillées ont ensuite servi de base à l'estimation des activités EMGs (Fig2, cadre du bas) : pour chaque muscle, l'amplitude de ces enveloppes a été normalisée par rapport à l'amplitude maximale observée pour ce muscle sur l'ensemble du corpus, permettant ainsi de comparer par la suite les activités des différents muscles en les exprimant en % des activités maximales observables sur chaque muscle. L'activité EMG sur une fenêtre temporelle d'intérêt a été définie comme l'intégrale de l'enveloppe redressée du signal EMG, divisée par la durée de la fenêtre. Nous avons ainsi mesuré l'activité EMG de chacun des cinq muscles lors des quatre phases (P1-P4) de chaque mouvement de production des occlusives /p/ ou /b/.

## 2.5 Analyses statistiques

Pour chaque muscle et chaque mouvement effectué, nous avons créé une variable réponse prenant comme valeur la phase dans laquelle l'activité musculaire est maximale. Nous considérons cette variable réponse comme une variable catégorielle ordonnée (4 niveaux : P1 à P4).
Nous avons testé l'influence du contexte vocalique (3 niveaux : /a/, /i/, /u/), du niveau d'effort (5 niveaux), du muscle (5 niveaux, DIG, DLI, MENT, OOI et OOS) et de leurs interactions sur cette variable réponse, à l'aide d'une régression ordinale avec effets aléatoires (package ordinal de R).
Nous avons utilisé une procédure de sélection pas à pas, utilisant des tests de rapport de vraisemblance, pour sélectionner les termes du modèle qui apportent de l'information significative.
Nous avons ensuite effectué des comparaisons multiples (package mulcomp de R) pour tester, pour chaque muscle, la significativité des différences observée entre les contextes vocaliques.

Enfin, des analyses de corrélation de Spearman ont été conduites pour examiner la corrélation entre l'activité de chacun des 5 muscles, dans chaque phase du mouvement, et les 3 descripteurs cinématiques du mouvement des lèvres (compression labiale, vitesses de fermeture et de réouverture).

# 3 Résultats

## 3.1 Patrons d'activation musculaire orofaciale lors de la production de consonnes occlusives – Influence de la voyelle adjacente

La figure 3 ci-dessus représente les activités musculaires moyennes (calculées sur toutes les consonnes labiales et tous les locuteurs) des 5 muscles d'intérêt, selon que la voyelle suivant la consonne est /a /, /i/, ou /u/. On note d'abord une importante similitude dans l'évolution de ces activités pour chacune des voyelles et pour l'ensemble des muscles à l'exception de l'OOS : l'activité EMG croit dès la phase P1 pour atteindre un pic très marqué dans la phase P3 avant de décroitre dans la phase P4. Le muscle OOS est lui essentiellement actif en phase P1 et P2, et devient très faiblement actif dès la phase P3, sauf quand la voyelle suivant la consonne est un /u/. Pour les voyelles non labiales /i/ et /a/ les muscles présentant les activités les plus fortes sont les muscles « ouvreurs » DIG, DLI et MENT. Lorsque la voyelle suivante est /u/, l'activité de ces trois muscles est sensiblement réduite et ce sont les muscles OOS et OOI qui semblent les plus activés.

Dans tous les cas on note la pertinence de la décomposition en 4 phases : les activités EMG de tous les muscles varient entre les phases P2 et P3, qui, rappelons-le, correspondent à l'occlusion labiale au cours de laquelle les lèvres, en contact, sont sensiblement immobiles (i.e. à vitesse quasi nulle).

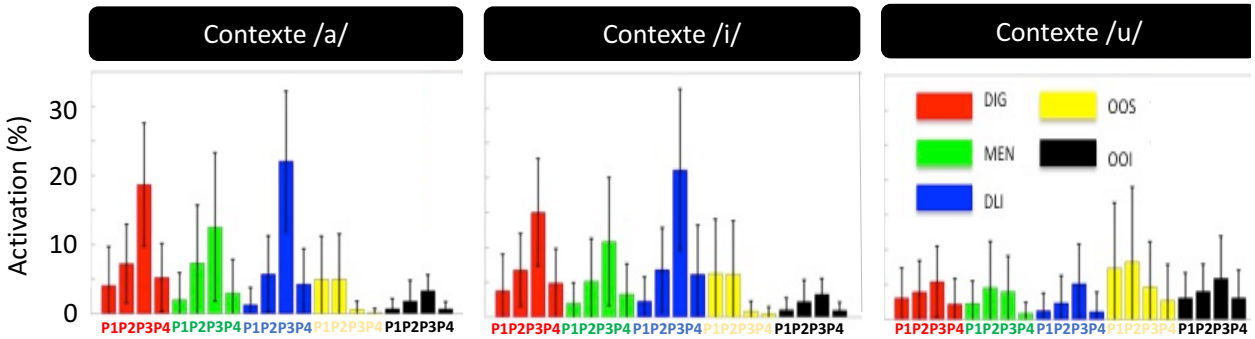



FIGURE 3 : Evolution de l'activité EMG (exprimée en pourcentage de l'activité maximum dans tout le corpus) au cours des phases P1 à P4, pour les cinq muscles étudiés, moyennée sur l'ensemble de consonnes labiales et sur les 4 sujets, lorsque la voyelle suivant la consonne est /a/, /i/ ou /u/.

L'analyse statistique de ces données selon le modèle décrit en section 2.5 permet de préciser ce premier constat global. La phase du mouvement dans laquelle une activité musculaire maximale est observée dépend significativement du muscle considéré, du contexte vocalique, du niveau d'effort, ainsi que de l'interaction entre le muscle considéré et le contexte vocalique ($chisq(8)=188.46$, $p<0.0001$), et de l'interaction entre le muscle considéré et le niveau d'effort ($chisq(16)=94.213$, $p<0.0001$). En contexte /a/ et /i/ les amplitudes des activités musculaires sont du même ordre de grandeur, tandis que l'amplitude des activités musculaires est plus faible pour le contexte vocalique /u/).

### 3.2 Corrélations entre l'activité musculaire et le déplacement des lèvres

Les analyses de corrélation entre l'activité de chacun des 5 muscles dans les 4 phases du mouvement, et les 3 descripteurs cinématiques du déplacement des lèvres (compression labiale, vitesse de fermeture, vitesse de réouverture) sont synthétisées sur la Figure 4. Celle-ci représente, pour chaque muscle et chaque phase du mouvement, la distribution des 24 coefficients de corrélation (Rho de Spearman) observés pour les 6 logatomes et les 4 locuteurs (chaque corrélation étant calculée sur environ 40 occurrences d'un même logatome). Cette représentation permet de repérer les cas pour lesquels une forte corrélation positive (>50%) est très fréquemment ou quasi-systématiquement observée entre l'activité d'un muscle dans une phase du mouvement, et l'un des 3 descripteurs cinématique du mouvement (distributions présentant un pic très marqué autour de 75%), et les cas pour lesquels l'activité musculaire n'est que peu corrélée, ou avec un degré de corrélation très variable, avec les descripteurs du mouvement.

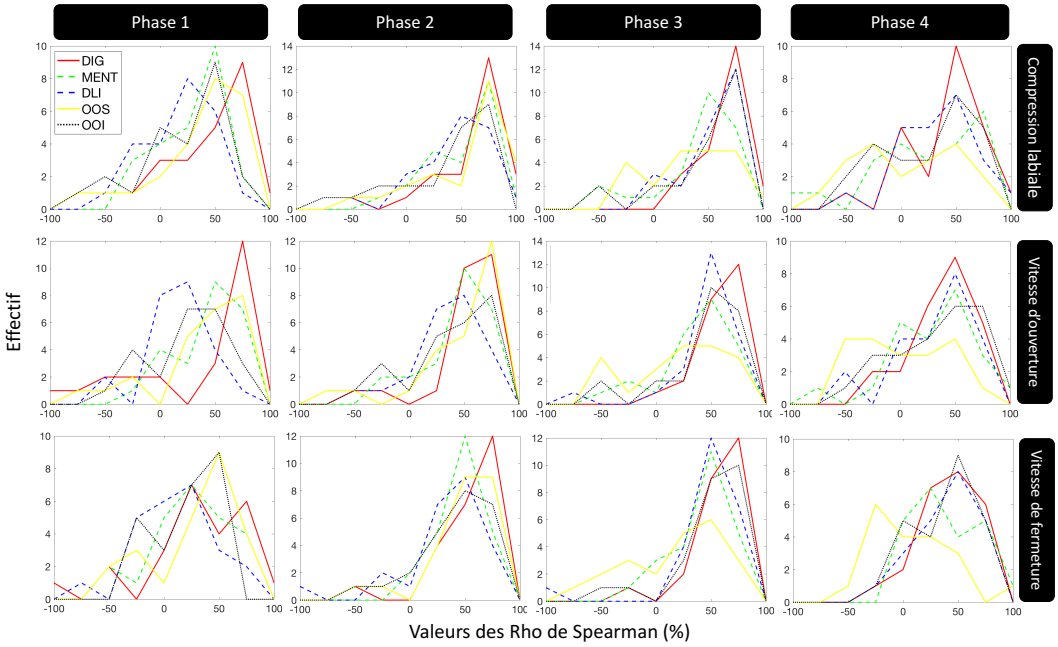



FIGURE 4 : Distribution des coefficients de corrélation (Rho de Spearman) entre l'activité de chacun des 5 muscles (DIG, MENT, DLI, OOS et OOI) dans les 4 phases du mouvement, et les 3 descripteurs cinématiques du déplacement des lèvres (compression labiale, vitesses de fermeture et de réouverture labiale).

Parmi les différentes corrélations d'intérêt, on note en particulier, que la compression labiale corrèle fortement, et de façon très reproductible, avec l'activité du MENT et de l'OOS en phase 2 du mouvement. On remarque également que la vitesse de ré-ouverture labiale corrèle fortement, et de façon très reproductible, avec l'activité du DIG en phase 3 du mouvement. Ces 3 activités musculaires repérées pourraient donc être de bons candidats comme descripteurs physiologiques de l'effort articulatoire. Ces pistes seront explorées plus en détails dans une prochaine base de données sur davantage de locuteurs.

# 4 Discussion

Pour les voyelles non labiales /i/ et /a/, l'analyse globale de la variation de l'activité des 5 muscles de la phase P1 à la phase P4 suggère que le muscle OOS (actif en phase P1 et P2) joue un rôle majeur dans la fermeture labiale, de l'initiation du mouvement jusqu'à la compression maximale, alors que le DIG, le DLI et le MENT, qui ont une activité croissante jusqu'en phase P3

jouent un rôle majeur dans le geste de réouverture. Ces observations sont cohérentes pour les muscles OOS, DIG et DLI avec les connaissances générales sur le rôle fonctionnel de ces différents muscles. En revanche, elles sont surprenantes pour le MENT qui est plutôt considéré comme un muscle élévateur de la lèvre. Deux explications potentielles à cette contradiction peuvent être formulées. La première serait que les signaux EMG collectés sur l'électrode du MENT soient en fait contaminés par les potentiels d'action des muscles abaisseurs adjacents, en particulier le DLI (phénomène de diaphonie). La seconde serait qu'il existe un mécanisme de coactivation des muscles élévateurs et abaisseurs de la lèvre pour garantir une meilleure stabilité du geste d'ouverture, dont on sait qu'il est crucial pour la production des plosives. Cette hypothèse serait cohérente avec le concept de « synergie » entre muscles antagonistes proposé par Latash (2007), observé de manière quasi-systématique dans les mouvements du bras, et qui permettrait de stabiliser les mouvements.

Il est intéressant de noter que pour la voyelle /u/ les muscles les plus actifs sont le OOS et le OOI qui jouent un rôle majeur (cf par exemple Nazari et al., 2011) sur l'arrondissement des lèvres, crucial pour la voyelle /u/. L'activation plus forte de ces deux muscles débute bien avant le geste de la consonne vers la voyelle. On note donc au niveau musculaire l'effet d'anticipation du geste d'arrondissement vocalique dans la consonne, déjà largement observé sur le plan cinématique (Noiray et al., 2011). On note aussi, que pour cette voyelle, l'OOS est actif en phase P3, ce qui est cohérent avec le rôle que nous attribuons à son activation pour la production de l'arrondissement vocalique.

De manière générale, il est important de noter qu'aucun des muscles considérés dans cette étude n'est actif en phase P4, qui est la phase d'ouverture où se produisent le bruit de plosion et les transitions formantiques vers la voyelle, dont la littérature a montré qu'ils sont des facteurs cruciaux de l'intelligibilité consonantique. Cela suggère que les muscles responsables du mouvement des lèvres lors de la production de la consonne ne conduisent pas le mouvement d'ouverture, mais déterminent les caractéristiques dynamiques des lèvres pour que leur vitesse d'ouverture et leur rigidité permettent les bonnes interactions entre les tissus et l'air pour donner les phénomènes aérodynamiques nécessaires à la réalisation de la consonne.

# 5 Conclusion

Les données présentées dans cet article ont permis de :

1. décrire temporellement l'activation musculaire de chaque muscle péri-oral ;
2. montrer l'intérêt d'une analyse temporelle plus détaillée de l'activité musculaire lors de la production de parole, ici de consonnes occlusives ;
3. identifier des descripteurs physiologiques de l'effort articulatoire, reproductibles entre les locuteurs et les segments de parole produits. Ainsi, trois descripteurs parmi d'autres possibles s'avèrent pertinents: l'activité des muscles MENT et OOS au cours de la phase de compression labiale (P2), et l'activité du muscle DLI au cours de la phase de relâchement de la compression (P3).

# Remerciements

A Thomas Hueber de Gipsa-lab pour ses conseils avisés lors de la réalisation du programme de détection des contours labiaux.

Cette recherche est financée par l'Agence Nationale de la Recherche (Projet StopNCo : Effort et coordination dans la production des consonnes occlusives ; ANR-14-CE30-0017 ; Maëva Garnier).

# Références


BECK T.W, HOUSH T.J, CRAMER J.T, WEIR J.P (2008). The effects of electrode placement and innervation zone location on the electromyographic amplitude and mean power frequency versus isometric torque relationships for the vastus lateralis muscle. *Journal of Electromyography and Kinesiology,* 18, 317-328.

BLAIR C, SMITH A (1986). EMG recording in human lip muscles : can single muscles be isolated ? *Journal of Speech and Hearing Research,* 29, 256-266.

CAMPANINI I, MERLO A, DEGOLA P, MERLETTI R, VEZZOZI G, FARINA D. (2007). Effect of electrode location on EMG signal envelope in leg muscles during gait. *Journal of Electromyography and Kinesiology,* 17, 515-526.

HOGREL J.-Y, DUCHÊNE J, MARINI J.-F (1998). Variability of some SEMG parameter estimates with electrode location. *Journal of Electromyography and Kinesiology,* 8, 305-315.

JENSEN C, VASSELJEN O, WESTGAARD R.H, (1993). The influence of electrode position on bipolar surface electromyogram recordings of the upper trapezius muscle. *European Journal of Applied Physiology,* 67, 266-273.

LALLOUACHE, M. T. (1991). *Un poste «Visage-parole» couleur. Acquisition et traitement automatique des contours des lèvres*. Thèse de Doctorat, Institut National Polytechnique de Grenoble, Grenoble, France.

LAPATKI B.G, OOSTENVELD R, VAN DIJK J.P, JONAS I.E, ZWARTS M.J, STEGEMAN D.F (2010). Optimal placement of bipolar surfaceEMGelectrodes in the face based on single motor unit analysis. *Psychophysiology,* 47, 299-314.

LATASH M.L, SCHOLZ J.P, SCHÖNER G. (2007). Toward a New Theory of Motor Synergies. *Motor Control*, 11, 276-308.

MCCLEAN M.D, TASKO S.M, (2003). Association of Orofacial Muscle Activity and Movement During Changes in Speech Rate and Intensity. *Journal of Speech, Language, and Hearing Research,* 46, 1387-1400.

NAZARI, M. A., PERRIER, P., CHABANAS, M., PAYAN, Y. (2011). Shaping by stiffening: a modeling study for lips. *Motor control*, 15(1), 141-168.

NOIRAY, A., CATHIARD, M. A., MÉNARD, L., ABRY, C. (2011). Test of the movement expansion model: Anticipatory vowel lip protrusion and constriction in French and English speakers. *The Journal of the Acoustical Society of America,* 129(1), 340-349.

O'DWYER N.J, QUINN P.T, GUITAR B.E., ANDREWS G, NEILSON P.D (1981). Procedures for verification . of electrode placement in EMG studies of orofacial and mandibular muscles. *Journal of Speech and Hearing Research,* 24, 273-288.

ROY S.H, DE LUCA C.J, SCHNEIDER J, (1986). Effects of electrode location on myoelectric conduction velocity and median frequency estimates. *The American Physiological Society,* 67, 1510-1517.

# L'incidence de la correction phonétique sur l'acquisition des voyelles en langue étrangère : étude de cas d'anglophones apprenant le français.

Charlotte Alazard-Guiu[1], Fabian Santiago[2], Paolo Mairano[3]
(1) U.R.I Octogone-Lordat, Université de Toulouse, 31058 Toulouse, France
(2) UMR 7023, SFL, Université Paris 8, 75017, Paris, France
(3) LFSAG, Université de Turin, 10100, Turin, Italie
charlotte.alazard@univ-tlse2.fr, fabian.santiago-vargas@univ-paris8.fr, paolo.mairano@unito.it

## RÉSUMÉ

Cet article a pour objectif d'étudier l'incidence de la correction phonétique sur l'acquisition des voyelles en L2. Des apprenants ont suivi des cours de correction phonétique selon deux méthodes différentes, la Méthode Verbo-Tonale et la Méthode Articulatoire. Dans une précédente étude (Alazard, 2013) nous avons comparé l'impact de chacune de ces méthodes sur le développement de la fluence. Dans cette étude, nous avons voulu tester l'incidence de la méthode utilisée sur les valeurs formantiques (F1-F2-F3) des voyelles produites par les apprenants avant et après les deux types d'entraînement. Nos premiers résultats montrent un effet du temps sur les valeurs formantiques de F3 mais pas d'effet de la méthode. Cette étude, bien que préliminaire, nous amène à penser que l'acquisition des voyelles en L2 met en jeu des processus spécifiques associés à une approche bimodale de la parole, que l'entrainement soit ou non explicitement focalisé sur les mouvements des lèvres.

## ABSTRACT

**The incidence of phonetic correction on the acquisition of L2 vowels : a case study of L1 English learners of L2 French.**

This article aims at studying the incidence of phonetic correction on the acquisition of L2 segmental patterns. Foreign L2 French learners attended phonetic classes with two different methods: the Verbo-Tonal vs. the Articulatory method. In a previous study (Alazard, 2013), we compared the incidence of these two methods on the acquisition of L2 fluency. In this study, we measured vowel formants (F1-F2-F3) on learners' productions before and after training. Our first results show an effect of time on F3 but no effect of the method. This study, though preliminary, leads us to think that vowel acquisition implies specific processes associated with a bimodal approach of speech, even if the training does not explicitly focus on lip movements.

MOTS-CLÉS : Acquisition d'une langue étrangère, Phonétique corrective, Formants vocaliques

KEYWORDS: Second Language Acquisition, Phonetic correction, Vowel formants


10.21437/JEP.2018-14

# 1 Introduction

Si l'oral a aujourd'hui une place reconnue dans l'enseignement du Français Langue Etrangère (désormais FLE), la place laissée à la phonétique corrective ne va pas de soi. On préfère généralement substituer à l'oral des quasi-synonymes, tels qu'interaction ou communication, qui ne rendent pas compte de sa nature même, et en particulier de sa dimension sonore. Ce constat peut en partie s'expliquer par l'existence de présupposés tenaces, notamment l'idée selon laquelle il n'est plus possible de maîtriser les aspects sonores d'une L2 après une période donnée ou encore l'idée selon laquelle la prononciation s'acquiert naturellement par des contacts répétés avec la langue cible.

Pourtant, des études montrent que des apprenants peuvent acquérir une prononciation perçue comme native grâce à un entrainement intensif en perception et/ou en production des sons de la langue étrangère (Bongaerts *et al*, 2000 ; Birdsong, 2003). En outre, dans une étude longitudinale (Alazard, 2013), nous avons comparé l'impact de deux méthodes de correction phonétique sur l'acquisition de la fluence en L2Nous avions retenu deux méthodes qui offraient deux approches opposées de l'apprentissage de la prononciation : une approche sans métalangage et plus 'implicite' avec la méthode verbo-tonale (MVT) et une approche descriptive et explicite avec la méthode articulatoire (ART).

La MVT se base en effet sur une approche procéduralisée de l'enseignement de la prononciation. Cette méthode utilise la structure prosodique de la langue cible comme support à l'acquisition de compétences phonologiques en L2. Plus spécifiquement, les patrons rythmiques de la langue cible sont utilisés pour mettre en lumière les spécificités de la langue cible (accentuation, intonation et phonèmes). L'enseignant va aider l'apprenant à se familiariser avec la structure prosodique de la langue cible à travers la répétition de logatomes (tels que « lalala » ou « dadada ») ou l'utilisation de gestes facilitateurs (par exemple un battement de la main pour souligner les proéminences). Dans une seconde phase, la structure prosodique et la gestualité accompagnatrice sont utilisées pour faciliter la perception et la reproduction des phonèmes. Par exemple, si l'apprenant assombrit le timbre du phonème cible, l'enseignant prononcera ledit phonème dans un contexte prosodique éclaircissant (c'est-à-dire à l'intérieur d'une syllabe accentuée), avec un geste de tension et demandera à l'apprenant de répéter le son dans ce nouveau contexte (Billières, 2005). La MVT se focalise sur un apprentissage non-explicite de la prononciation aidé par la gestuelle.

À l'opposé, la méthode articulatoire, postule qu'une bonne production implique la connaissance métalinguistique de l'articulation des sons, autrement dit, la mise en place de connaissances déclaratives. En conséquence, l'enseignant donne une description articulatoire des différents sons puis invite l'apprenant à répéter le geste articulatoire afin de produire le son cible. Par exemple, pour produire un /y/ l'enseignant dira à l'apprenant de placer sa langue à l'avant de sa bouche et d'arrondir les lèvres. Dans cette méthode, on met l'emphase sur la production et la répétition de sons isolés, puis de mots isolés contenant le son cible et finalement de phrases. Les paramètres prosodiques sont globalement négligés bien qu'une description métalinguistique des caractéristiques intonatives de la langue cible puisse être envisagée dans les pratiques de classes, le plus souvent en fin de formation.

Dans une précédente étude (Alazard 2013), nous avions montré que seul un entrainement à la correction phonétique par la MVT permettait à des apprenants débutants de développer une réelle compétence en fluence en L2, en seulement trois semaines. En effet, nous avions observé une diminution du nombre de pauses agrammaticales, une augmentation du taux d'articulation et une

augmentation du débit de parole uniquement pour les apprenants qui avaient suivi ces cours de correction phonétique. De plus, nous avions observé, chez ces mêmes apprenants, une nette diminution de la durée des syllabes inaccentuées et une augmentation de la durée des syllabes accentuées entre le premier et le deuxième test, indiquant une augmentation de la densité accentuelle et du contraste accentuées/inaccentuées. Ces variations de durée accentuelles semblaient indiquer que les apprenants produisaient un schéma accentuel plus proche de celui du français après entrainement. De même, ils étaient également capables de produire des groupes rythmiques beaucoup plus longs. Cela nous avait amené à penser que les apprenants qui avaient suivi un entrainement via la MVT avaient commencé à mettre en place des stratégies d'encodage adaptées à la L2 via la maitrise des patrons prosodiques du français.

Dans cette étude, nous souhaitons mesurer les bénéfices de ces deux types d'entraînement, non plus sur les aspects prosodiques, mais sur la prononciation des voyelles. Nous pensons que la MVT pourrait avoir une plus grande incidence sur la différenciation des timbres vocaliques des voyelles en position accentuées, grâce à l'attention consacrée au niveau prosodique et à l'utilisation de procédés correctifs visant à modéliser l'aperture et le lieu d'articulation par des gestes de tension ou de relâchement.

# 2 Méthodologie

Pour comparer l'impact des deux méthodes et tester notre hypothèse, nous avons repris les données de l'étude précédente (Alazard 2013). Ces données avaient été recueillies via une étude longitudinale pilote, sur trois semaines, à raison de 2 séances de correction phonétique de 90 min par semaine. Tous les cours avaient été dispensés par le même enseignant, formé aux deux méthodes.

Nous avions enregistré huit apprenants anglophones de niveau débutant divisés en deux groupes selon la méthode suivie (ART vs MVT). Le niveau des apprenants avait été évalué pour l'étude sur la base d'un entretien oral semi-directif selon l'échelle proposée par le CECRL. Ce test de niveau, inspiré des épreuves d'évaluation de la production et de la compréhension orales du DELF–DALF, comprenait plusieurs activités : un entretien guidé, des répétitions de phrases et une activité de compréhension orale. Les participants étaient âgés de 20 à 60 ans (moyenne d'âge 32,5 ans). Tous les participants résidaient en France au moment de l'étude. Ils avaient tous préalablement suivi un enseignement en français, soit au cours de leur scolarité dans leur pays d'origine, soit à leur arrivée en France. En parallèle des cours proposés dans le cadre de cette recherche, la majorité des apprenants suivaient également des cours de français.

Les apprenants avaient été testés dans une tâche de lecture oralisée à deux reprises, avant la formation (T1) et après trois semaines d'entrainement exclusivement à l'oral (T2). Les textes concernaient des passages/courtes histoires adaptés au niveau des apprenants. Les sujets avaient passé les tests individuellement dans une salle d'enregistrement adaptée, à l'université de Toulouse 2. Ils avaient pour consigne de s'approprier le texte par une ou plusieurs lectures silencieuses avant d'être enregistrés. Après avoir pris connaissance du texte, les sujets lisaient une fois le texte à voix haute avant de répondre à des questions de compréhension posées par l'examinateur.

Les données recueillies avaient d'abord été transcrites phrase par phrase selon les conventions de la Transcription Orthographique Enrichie (TOE) (Bertrand *et al*, 2008). Les phrases transcrites avaient ensuite été alignées automatiquement grâce au logiciel *SPPAS* (Bigi & Hirst, 2012) et

segmentées à différents niveaux : mots et syllabes. Pour cette étude, les voyelles ont été ré-étiquetées et réalignées manuellement dans *Praat* (Boersma & Weenink, 2005).

# 3 Annotation et analyse formantique des voyelles

## 3.1 Voyelles analysées

L'ensemble de notre corpus contenait 2704 segments vocaliques produits dans les tâches de lecture par les participants, dont 1361 voyelles produites avant la formation (T1), et 1340 après l'entraînement (T2). Nous avons utilisé une transcription canonique des mots, et non une transcription de la production réelle de l'apprenant (p. ex. pour le mot <déjeuner>, la transcription était [deʒœne], même si l'apprenant prononçait [dəʒune]). Cela nous a permis de vérifier dans quelle mesure les formants des voyelles ciblées changeaient (ou pas) en fonction de la méthode employée/après entrainement.

Nous avons évalué deux types de métriques : la première consistait à analyser le timbre des voyelles [i], [y] [e], [ø], [ɛ], [œ], [a], [o], [ɔ], [u] en fonction de leurs valeurs de F1, F2 et F3 de manière séparée. Notre objectif étant de vérifier dans quelle mesure les corrélats d'aperture, de position de la masse de la langue et des mouvements des lèvres étaient modifiés après les deux types d'entraînement. En outre, nous avons choisi d'évaluer les distances euclidiennes mesurées dans la charte de l'espace vocalique F2*F3 pour les voyelles [u]/[y], [e]/[ø] et [ɛ]/[œ] afin de mesurer le degré de réalisation d'opposition entre ces voyelles [les voyelles antérieures labialisées sont notamment marquées (Wurzel 1998) et posent problème aux apprenants anglophones (Darcy *et al.* 2012)]. Les voyelles produites en contexte de disfluence, faux départs, hésitations ou mots inachevés ont été exclues de l'analyse. Le nombre final de segments analysés s'élevait donc à 2,1k voyelles.

## 3.2 Extraction des valeurs formantiques

Les valeurs spectrales des voyelles (F1-F2-F3) ont été extraites dans la zone médiane de chaque voyelle via un script *Praat* afin d'obtenir des valeurs formantiques plus ou moins stables. L'amplitude des pics a été détectée avec une bande inférieure à 5 kHz pour les hommes et inférieure à 6 kHz pour les femmes avec une fenêtre Gaussienne de 25ms. Dans le but d'exclure les valeurs formantiques aberrantes dues à la mauvaise détection de formants (voix craquée, assourdissements, globalisations *etc.*), nous avons employé un filtre selon Gendrot & Decker (2007). Enfin, nous avons converti les valeurs formantiques Hz en Barks afin d'obtenir une normalisation par genre et par locuteur (Labov 2006)

## 3.3 Distances euclidiennes

A partir des valeurs formantiques obtenues en Barks dans la section précédente, nous avons calculé les distances euclidiennes sur le plan F2*F3 pour les paires de voyelles suivantes : [u]/[y], [e]/[ø] et [ɛ]/[œ]. Notre hypothèse était la suivante : une distance plus longue entre ces paires de voyelles sur le plan cartésien F2*F3 serait le résultat d'une tendance à catégoriser les voyelles en question, et, pour les paires [e]/[ø] et [ɛ]/[œ], traduirait plus particulièrement l'incidence des entraînements sur le trait de labialité. Autrement dit, si les participants bénéficient des avantages de l'une (ou des deux) méthodes, cela devrait avoir une incidence dans la position qu'occupent ces

voyelles dans leur espace acoustique : plus elles sont éloignées dans la charte vocalique l'une par rapport à l'autre, plus cela marque une tendance à les opposer.

# 4. Résultats

Pour notre étude, nous avons modélisé les résultats obtenus moyennant des modèles mixtes avec le package *lme4* sous R. La puissance des variables prédictives (groupe, temps et voyelles) sur les variables dépendantes (F1, F2, F3, ainsi que les distances euclidiennes en question) a été estimée moyennant des tests du rapport de vraisemblance (*likelihood ratio tests*).

## 4.1 Effets de la technique sur les valeurs formantiques

Nous analysons ici les valeurs formantiques obtenues (F1, F2 et F3) avant et après (T1 et T2 respectivement) dans les deux groupes (ART= articulatoire vs MVT = verbo-tonale). Les figures ci-dessous illustrent les espaces vocaliques sur le plan F1*F2 (Figure 1) et F2*F3 (Figure 2) selon les moyennes obtenues avant et après les traitements dans chacune des méthodes.

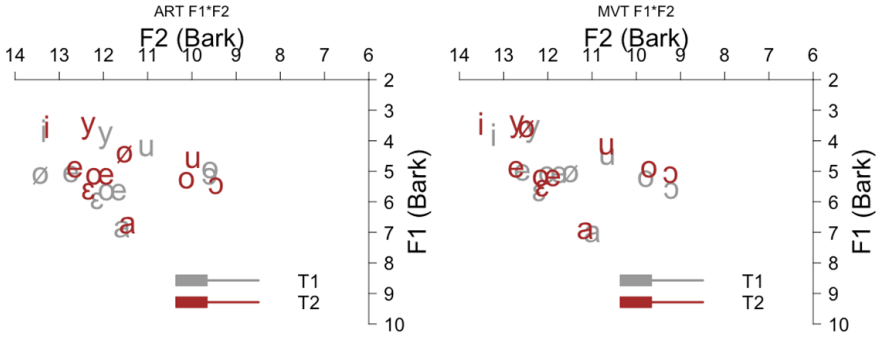



FIGURE 1 : Comparaison des triangles vocaliques sur le plan F1*F2 avant (T1) et après entraînement (T2) selon chacune des méthodes (Articulatoire vs Verbo Tonale).

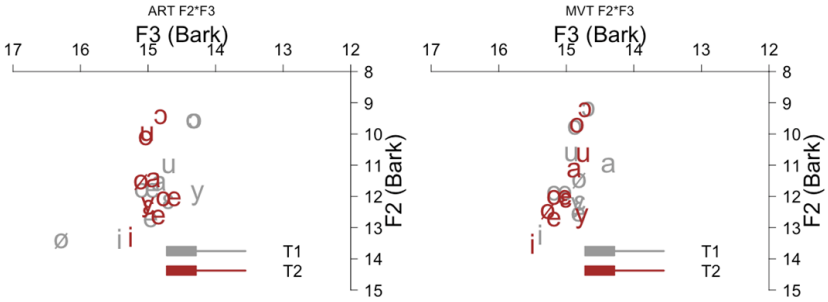



FIGURE 2 : Comparaison des triangles vocaliques sur le plan F2*F3 avant (T1) et après entraînement (T2) selon chacune des méthodes (Articulatoire vs Verbo Tonale).

Une inspection visuelle de ces espaces acoustiques ne nous permet pas de voir clairement les effets de chacune des méthodes, les voyelles se superposant entre elles dans la plupart des cas. Afin d'analyser l'impact des méthodes dans les valeurs formantiques des voyelles, nous avons construit plusieurs modèles (régressions linéaires à effets mixtes) avec « F1 », « F2 » et « F3 » comme variables dépendantes, « Groupe » (ART vs MVT), « Temps » (T1 vs T2), « Voyelle » (le phonème en question) comme facteurs fixes (ainsi que leurs possibles interactions) et

« Participants » comme variable aléatoire. Les résultats du modèle évaluant les effets du Groupe (T2 exclu) sur les trois variables dépendantes (F1, F2 et F3) n'ont pas atteint le seuil de significativité (toutes les valeurs de $p > 0,05$), ce qui indique que les deux groupes étaient homogènes avant l'entrainement et donc comparables dans leur niveau de maîtrise phonique. En revanche, nous n'avons pas trouvé d'effet de Groupe, ni de Temps sur les valeurs formantiques de F1 et F2 (toutes les valeurs de $p > 0,05$). Les résultats montrent que seules les valeurs de F3 sont affectées par le facteur Temps : elles sont plus élevées au T2 par rapport au T1 ($\beta = 0.193$, e. t. = 0,06, $t = 2,88$, $p < 0,01$), mais elles ne changent pas en fonction du Groupe ($p > 0,05$). En d'autres termes, les participants montrent une tendance à produire des valeurs de F3 plus élevées au T2 quelle que soit la technique employée. Tous ces résultants montrent qu'à l'exception de F3, les différences observées dans la figure ci-dessus sont certainement dues à la variabilité des participants, et qu'aucune des deux méthodes n'a d'incidence significative sur les valeurs formantiques de F1 et F2. Une analyse post-hoc par paires avec une correction de Tukey a montré que la voyelle [a] est en réalité la seule pour laquelle F3 est significativement différente entre T1 et T2 ($\beta = 0,34$, é.-t. = 0,09, $z = 3,69$, $p < 0,05$). Une analyse individuelle par participant a montré une importante variabilité interlocuteur pour toutes les voyelles et a révélé que les valeurs de F3 peuvent être affectées différemment en fonction du temps. Ainsi, seulement 3 participants sur 8 (2 du groupe ART et 1 du groupe MTV) produisent des valeurs de F3 plus basses au T2 pour la voyelle [œ] montrant ainsi que cette voyelle tend à être articulé avec un arrondissement des lèvres plus marqué après l'entrainement. Le reste des participants produisaient des valeurs de F3 plus hautes ou similaires. L'échantillon relativement petit de locuteurs analysés (4 par groupe) ne permet pas de tirer des conclusions définitives et plus de données seraient nécessaires pour une analyse statistique plus fiable.

## 4.2 Effets de la technique sur les distances euclidiennes

Notre deuxième objectif était de confirmer si les participants arrivaient à catégoriser les paires de phonèmes suivants [u]/[y], [e]/[ø] et [ɛ]/[œ] en termes des distances euclidiennes dans l'espace F2*F3 (plus précisément, l'emploi de la dynamique des lèvres combinée à une articulation vocalique antérieure pour opposer ces phonèmes). Notre hypothèse était que, si des effets positifs de la méthode employée étaient observés, les participants devraient articuler ces segments avec une plus grande distance acoustique. La figure ci-dessous illustre les distances euclidiennes (en Barks) entre les phonèmes en question, avant et après entrainement, en fonction de la méthode. Cette figure illustre qu'il y a une certaine tendance à marquer des distances plus longues pour les phonèmes [u]/[y], particulièrement avec la technique Articulatoire. En revanche, les résultats du modèle statistique montrent que ces différences ne sont pas significatives ($p > 0,05$). Cela suggère que ni le temps ni la méthode n'ont d'effet sur les distances euclidiennes entre ces paires de phonèmes produits par nos participants. Mais, encore une fois, l'échantillon limité de participants pourrait ne pas être en mesure de relever des différences.

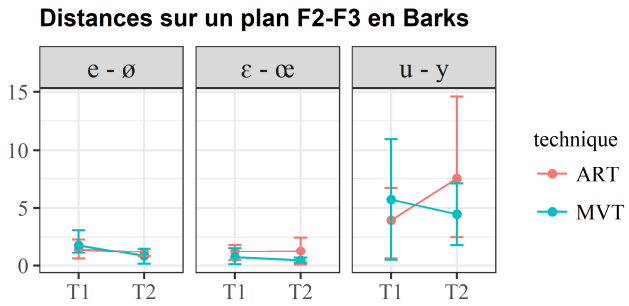



FIGURE 3 : Comparaison des triangles vocaliques sur le plan F2*F3 avant (T1) et après entraînement (T2) selon chacune des méthodes (Articulatoire vs Verbo Tonale).

# 5 Discussion

Dans cette étude, nous avons voulu comparer l'impact de deux méthodes de correction phonétique sur le développement du niveau segmental en L2. Nous pensions, comme pour la fluence dans une étude précédente, observer des différences entre les deux méthodes correctives choisies sur l'acquisition des voyelles. Nous avions formulé l'hypothèse selon laquelle la MVT permettrait une meilleure acquisition de F1 et F2. Pour tester cette hypothèse, nous avons analysé les valeurs F1, F2 et F3 de 2,1k segments vocaliques produits dans une tâche de lecture oralisée, avant et après entraînement à la correction phonétique, selon deux méthodes (Articulatoire vs. Verbo-Tonale). A partir des valeurs formantiques obtenues, nous avons également mesuré les distances euclidiennes sur le plan F2*F3 pour les paires de voyelles suivantes : [u]/[y], [e]/[ø] et [ɛ]/[œ], afin de déterminer si les apprenants avaient plus de facilités à catégoriser les voyelles données, avant et après entraînement, en fonction de la méthode.

Nos premiers résultats ne montrent pas d'incidence de l'apprentissage ni de la méthode sur les distances acoustiques ni sur les valeurs formantiques de F1 et F2. Les variations formantiques observées pourraient simplement être le résultat de la variabilité des participants. En revanche, nos résultats illustrent un effet du temps sur les valeurs formantiques de F3 pour la voyelle [a] chez tous les participants, après entrainement, quelle que soit la méthode utilisée. Les résultats issus des distances euclidiennes ne nous permettent cependant pas de dire si les participants arrivent à distinguer les paires de voyelles arrondies/non arrondies dans l'espace acoustique.

Une inspection individuelle par participant montre, par contre, que certains participants arrivent à baisser les valeurs formantiques de F3, indiquant ainsi que l'articulation de la voyelle [œ] est plus arrondie. Même si cela ne peut pas être généralisable, ces observations préliminaires nous amènent à penser que la correction phonétique des voyelles pourrait impliquer, que l'on en est conscience ou non, la prise en compte des mouvements des lèvres, faisant écho aux travaux sur le lien entre informations visuelles et information auditives en perception de la parole (voir par exemple : McGurk & MacDonald 1976 ; Skipper et al, 2007 ; Yeung & Werker, 2013). C'est d'autant plus intéressant que si la méthode articulatoire se focalise effectivement sur une description articulatoire des mouvements des différents articulateurs, la MVT utilise en revanche la gestualité accompagnatrice pour illustrer les mouvements articulatoires. D'après nos premiers résultats, il semblerait que la MVT ait des effets plus rapides dans l'amélioration de la composante prosodique que sur le plan segmental. En effet, en comparant les résultats positifs obtenus par Alazard (2013) sur la fluence avec les résultats reportés dans cette étude, il semblerait que l'acquisition des

voyelles nécessite un entrainement plus long. Afin de confirmer ces premières tendances, nous envisageons, comme dans l'étude précédente de comparer les résultats acoustiques obtenus avec un test sur la base de jugements auditifs effectué auprès d'enseignants de FLE. Ces données préliminaires nous amènent également à réfléchir à la corporalité accompagnatrice non plus comme quelque chose de défini a priori en fonction de la méthode choisie mais comme quelque chose à adapter en fonction des phénomènes que l'on veut faire acquérir aux apprenants de L2. De nouvelles études sont donc nécessaires pour d'une part confirmer les tendances observées, en élargissant notamment l'analyse à l'acquisition des consonnes, et d'autre part, s'interroger sur l'impact de la gestualité (micro ou macro) sur l'acquisition de la phonétique en L2, en regardant aussi les effets obtenus avec d'autres méthodes de correction phonétique.

# 5 Références


ALAZARD, C. (2013). Rôle de la prosodie dans la fluence en lecture oralisée chez des apprenants de Français Langue Etrangère. Thèse de doctorat, Université de Toulouse 2.

BERTRAND, R., BLACHE, P., ESPESSER, R., FERRE, G., MEUNIER, C., PRIEGO-VALVERBE, B. & RAUZY, S. (2008). Le CID-Corpus of Interactional Data-Annotation et exploitation multimodale de parole conversationnelle. *Traitement automatique des langues*, *49*(3), 1-30.

BIGI, B. & HIRST, D. (2012). Speech phonetization alignment and syllabification (SPPAS): a tool for the automatic analysis of speech prosody. *Paper presented at the 6th Speech Prosody Conference, Shanghai.*

BILLIERES, M. (2005). Les pratiques du verbo-tonal. Retour aux sources. In *Berré, M.(Éd.), Linguistique de la parole et apprentissage des langues. Questions autour de la méthode verbo-tonale de P. Guberina. Mons: CIPA*, 67-87.

BIRDSONG, D. (2003). Authenticité de prononciation en français L2 chez des apprenants tardifs anglophones : analyses segmentales et globales. *Acquisition et interaction en langue étrangère* 18, 17-36.

BOERSMA, P., & WEENINK, D. (2005). Praat: doing phonetics by computer, http://www.praat.org

BONGAERTS, T., MENNEN, S., & SLIK, F.V.D. (2000). Authenticity of pronunciation in naturalistic second language acquisition: The case of very advanced late learners of Dutch as a second language. *Studia linguistica* 54, 298-308.

DARCY, I., DEKYDTSPOTTER, L., SPROUSE, R. A., GLOVER, J., KADEN, C., MCGUIRE, M., & SCOTT, J. H. (2012). Direct mapping of acoustics to phonology: On the lexical encoding of front rounded vowels in L1 English–L2 French acquisition. *Second Language Research*, *28*(1), 5-40.

GENDROT, C. & ADDA-DECKER, M. (2007). Impact of duration and vowel inventory siez on formant values of oral vowels: an automated formant analysis from eight languages. Actes de *16th International Conference of Phonetic Sciences,* 1417-1420.

LABOV, W. (2006). A sociolinguistic perspective on sociophonetic research. *Journal of Phonetics*, 34, 500-515.

MCGURK, H., & MACDONALD, J. (1976). Hearing lips and seeing voices. *Nature* 264 (5588), 746-748.

SKIPPER, J. I., VAN WASSENHOVE, V., NUSBAUM, H. C., & SMALL, S. L. (2007). Hearing lips and seeing voices : how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral cortex* 17 (10), 2387-2399.

YEUNG, H. H., & WERKER, J. F. (2013). Lip movements affect infant's audiovisual speech perception. *Psychological Science* 24 (5), 603-612.

WURZEL, W. U. (1998). On markedness. *Theoretical Linguistics*, *24* (1), 53-72.

# Ambiguïté temporaire des obstruantes voisées en parole chuchotée

Yohann Meynadier, Sophie Dufour
Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
yohann.meynadier@univ-amu.fr, sophie.dufour@lpl-aix.fr

## RESUME

Meynadier & Dufour (2016) ont étudié la résolution perceptive d'ambiguïtés de voisement en parole chuchotée en français. Via un paradigme d'amorçage sémantique, des mots cibles visuels ont été présentés immédiatement après la fin de mots amorces audio sémantiquement reliées aux cibles. En parole chuchotée, un effet d'amorçage a été observé seulement pour les amorces contenant une obstruante sourde (/pe*t*al/ amorce FLEUR), mais pas pour les amorces à obstruante voisée : /pe*d*al/ ne facilitait ni le traitement de son associé sémantique VÉLO, ni celui de FLEUR, relié à l'autre membre de la paire minimale à obstruante sourde (/pe*t*al/). Dans cette étude, les cibles visuelles sont présentées 50 ms après les amorces audio. Contrairement à ce qui a été observé avec un délai de 0 ms, seules les mots chuchotés à obstruante voisée amorcent leur associé sémantique (/pe*d*al/–VÉLO), ce qui n'est plus le cas des amorces chuchotées à obstruante sourde (/pe*t*al/–FLEUR). Ces études suggèrent que des indices résiduels du voisement sont exploités par les auditeurs pour retrouver le mot chuchoté. Néanmoins, la reconnaissance du trait [+voisé] en parole chuchotée ne serait pas immédiate et nécessiterait un certain temps supplémentaire de traitement.

## ABSTRACT

**Temporary ambiguity of voiced obstruantes in whispered speech**
Meynadier & Dufour (2016) have studied how French listeners resolve voicing ambiguities in whispered speech. Visual targets were presented at the offset of auditory semantically-associated primes. A similar priming effect to that observed with modal primes was found only with whispered primes included a voiceless consonant (/pe*t*al/ *"pe**t**al"* primed FLEUR *"flower"*). No priming effect was found with whispered primes included a voiced consonant (/pe*d*al/ *"pe**d**al"*), either on the target VÉLO "bike" related to /pe*d*al/ or on the target FLEUR related to /pe*t*al/. Here, visual targets were presented 50 ms after the offset of primes. While priming effects were no longer observed with whispered voiceless primes (/pe*t*al/–FLEUR), a priming effect emerged with whispered voiced primes (/pe*d*al/–VÉLO). Together, our results suggest that residual correlates of voicing present in whispered speech are exploited by listeners to recover the intended words. Nonetheless, they also showed that the reconstruction of the voiced feature is not immediate during whispered word recognition and requires a certain amount of time to be done.

MOTS-CLES : voisement, parole chuchotée, phonologie, reconnaissance lexicale
KEYWORDS: voicing, whispered speech, phonology, lexical recognition

# 1 Introduction

Cette étude constitue un second volet à nos travaux sur la reconnaissance des mots chuchotés en français, particulièrement focalisé sur l'ambiguïté potentielle des obstruantes voisées du fait de


10.21437/JEP.2018-15

l'absence totale de vibrations laryngées dans ce mode de phonation. Cette absence de vibrations laryngées provoque d'importantes modifications acoustiques qui affectent davantage les segments voisés que les segments sourds (Ito et al. 2005, Jovičić & Šarić 2008). Logiquement, la modification principale concerne le trait [+voisé] phonétiquement réalisé par un assourdissement complet des segments phonologiquement voisés. Ainsi, en français un mot tel que *pédale* /pe*d*al/ devient potentiellement ambigu parce que sa prononciation en voix chuchotée se rapproche de manière critique de celle de *pétale* /pe*t*al/, son seul voisin phonologique en paire minimale de voisement. Néanmoins, les assimilations complètes de voisement ne neutraliseraient pas totalement les oppositions par le seul trait de voisement. Ainsi, par exemple, Hallé et al. (2012) observent que les consonnes assimilées en voisement maintiennent des durées phonétiques propres à leur identité sous-jacente [±voisé], à savoir plus longues pour les [–voisé] que pour [+voisé]. Kohlberger & Strycharczuk (2015) rapportent un résultat similaire pour la consonne /s/ en contexte d'assimilation de voisement en parole chuchotée. Par ailleurs, quelques études montrent que des corrélats physiologiques du voisement sont préservés en voix chuchotée, tels que des différences de pression intraorale (Netsell 1969, Slis 1970, Garnier et al. 2014, Meynadier 2015) et d'ouverture glottique (Malécot & Peebles 1965, Mills 2009, Meynadier 2015). Sur le plan acoustique, alors que les mesures spectrales (par ex., transition du F1, intensité et fréquence de bruit) semblent peu ou variablement corrélées au contraste voisement (Tartter 1989, Jovičić & Šarić 2008, van de Velde & van Heuven 2011, Gilichinskaya 2012, Kohlberger & Strycharczuk 2015), les durées des consonnes et des voyelles préconsonantiques se révèlent y être plus systématiquement associées (Sharf 1964, Schwartz 1972, Parnell et al. 1977, Tartter 1989, Jovičić & Šarić 2008, Vercherand 2010, van de Velde & van Heuven 2011, Gilichinskaya 2012, Meynadier & Gaydina 2013), comme en voix modale.

Les quelques études ayant examiné la perception du voisement des obstruantes dans des mots, non-mots ou syllabes montrent qu'elle est assez bien maintenue (au-dessus du seuil du hasard) en voix chuchotée, bien que moins performante qu'en voix modale (Dannenbring 1980, Tartter 1989, Munro 1990, Mills 2009, Vercherand 2010, Fux 2012, Gilinchskaya 2012). Cependant, si l'on s'y attarde, pour la plupart, cette bonne performance repose en grande partie sur un biais perceptif en faveur des obstruantes [–voisé] montrant un taux de reconnaissance souvent bien supérieur à celui des obstruantes [+voisé]. Par exemple, dans les études sur l'anglais le pourcentage d'identification correcte se situe entre 81 et 87 % pour les [–voisé] et entre 68 et 80 % pour les [+voisé] (Tartter 1989, Gilinchskaya 2012). En français, Fux (2012) observe un taux de 80 à 87 % pour les [–voisé] et de 16 à 47 % pour les [+voisés]. Ce biais perceptif en faveur des [–voisé] atténuerait ainsi l'hypothèse de corrélats résiduels du trait de voisement parfaitement exploités par les auditeurs pour reconnaître les mots chuchotés présentant un contraste de voisement. Plus généralement, cela suggère que les auditeurs, et plus particulièrement les auditeurs français, échoueraient à reconnaitre des mots chuchotés à obstruante voisée, comme *pédale* /pe*d*al/, et pourraient les confondre avec leur voisin phonologique à obstruante sourde, tel que *pétale* /pe*t*al/. Une limite partagée par toutes ces études est l'utilisation de tâches off-line, sans contrainte de temps de réponse, le plus souvent à choix forcé, favorisant l'implication de processus métalinguistiques ne permettant pas de refléter précisément les processus en temps réel de reconnaissance des mots parlés.

Pour cette raison, nous avons jugé utile de réexaminer la perception du trait de voisement à l'aide d'un paradigme d'amorçage sémantique, connu pour refléter des processus en temps réel de la reconnaissance des mots parlés, et en particulier l'activation des représentations lexicales (Tabossi 1996). Relié à notre question, ce paradigme consiste à présenter auditivement un mot amorce chuchoté potentiellement ambigu suivi d'un mot cible présenté visuellement, sémantiquement relié ou non à l'amorce, sur lequel les participants ont à réaliser une tâche de décision lexicale. Dans une première étude (Meynadier & Dufour 2016), nous avions observé que seules des amorces chuchotées à obstruante [–voisé] (/pe*t*al/) facilitaient le traitement ultérieur d'un mot cible relié

sémantiquement (FLEUR), alors qu'aucun effet d'amorçage de la cible reliée sémantiquement (VÉLO) n'apparaissait pour les amorces chuchotées à obstruante [+voisé] (/pé*d*ale/). De plus, ni les amorces à obstruante [–voisé], ni celle à obstruante [+voisé] précédant la cible reliée à l'autre membre de la paire minimale ne montrait d'effet d'amorçage (/pe*t*al/-VÉLO ou /pe*d*al/-FLEUR). Un point clé dans cette étude était que les cibles visuelles étaient présentées immédiatement après la fin acoustique des amorces auditives, c'est-à-dire avec un délai de 0 ms. Ces résultats laissaient ainsi suggérer qu'au moment où les mots cibles étaient présentés, si les obstruantes sourdes n'étaient plus ambiguës quant à leur trait sous-jacent, l'ambigüité liée au trait [+voisé] n'était quant à elle pas encore résolue. En effet, /pe*d*al/ n'amorçait pas sa cible VÉLO, mais n'était pas non plus confondu avec /pe*t*al/, puisque /pe*d*al/ chuchoté n'amorçait pas non plus FLEUR (cible reliée sémantiquement à /pe*t*al/). Dès lors, il s'avérait qu'un certain temps supplémentaire de traitement pouvait être nécessaire à la reconstruction du trait de voisement en parole chuchotée.

Afin de tester cette hypothèse, nous avons reconduit ici exactement la même expérience que dans notre étude de 2016, excepté que le délai entre la fin de la présentation des amorces audio et des cibles visuelles était allongé à 50 ms. Nous nous attendons donc qu'avec ce délai de traitement rallongé les mots chuchotés à obstruante [+voisé] amorcent uniquement leur cible sémantiquement reliée. Dès lors, le trait [+voisé] pourrait bien être identifié, probablement sur la base des corrélats phonétiques du voisement préservés en phonation chuchotée, à savoir plus précisément des durées consonantiques et/ou de la voyelle précédente. Afin d'évaluer l'implication des corrélats spectraux et temporels dans la reconnaissance du trait de voisement des obstruantes chuchotées, une analyse acoustique des mots amorces est proposée, avant de conclure sur une discussion générale de ces résultats de perception et de production.

# 2 Perception des obstruantes chuchotées

Les expériences de cette étude reprennent exactement le même matériel linguistique, le même design expérimental et les mêmes stimuli enregistrés pour l'étude de Meynadier & Dufour (2016). Nous y renvoyons le lecteur pour plus de détails. Ici, nous rappellerons l'essentiel de ce protocole et des différences avec l'étude antérieure, avant de présenter et de discuter les nouveaux résultats.

## 2.1 Matériel linguistique et design expérimental

Toutes les oppositions distinctives de voisement du français ont été testées dans ces expériences : /p t k f s ʃ/ *vs* /b d g v z ʒ/. Ces consonnes en opposition apparaissaient en position intervocalique de 2 x 20 mots en paire minimale (par ex., pé*t*ale–pé*d*ale, des*s*ert–dé*s*ert, am*ph*i–en*v*ie). Ces 40 mots ont été soumis à un pré-test d'association sémantique libre passé par 8 hommes et 30 femmes (français, 20-47 ans), afin d'établir les couples amorce–cible reliée les plus fréquents (à savoir > 20% des réponses), par ex. pé*t*ale–FLEUR, pé*d*ale–VÉLO, des*s*ert–CHOCOLAT, dé*s*ert–SABLE, am*ph*i–COURS, en*v*ie–DÉSIR. Les caractéristiques de ces amorces et de ces cibles, telles que le taux d'association sémantique, le nombre de phonèmes, de lettres et de syllabes, ainsi que leur point d'unicité et leur fréquence cumulée, ont été contrôlées (cf. Table 1, Meynadier & Dufour 2016). Pour mesurer l'effet d'amorçage attendu, chaque cible d'une amorce en paire minimale a été associée à un mot non relié sémantiquement (par ex., quittance–FLEUR/VÉLO, jumelle–CHOCOLAT/SABLE, héros–COURS/DÉSIR), constituant ainsi les conditions de contrôle comparées aux conditions de test où l'amorce est issue d'une paire minimale de voisement et reliée sémantiquement à la cible. L'éventualité d'un amorçage croisé étant aussi testé, les amorces en paire minimale ont également été associées à la cible reliée à l'autre membre de la paire minimale en opposition de voisement, par ex. pé*t*ale–VÉLO, pé*d*ale–FLEUR, des*s*ert–SABLE, dé*s*ert–CHOCOLAT, am*ph*i–DÉSIR, en*v*ie–COURS. Les listes correspondant à ces huit conditions expérimentales sont rapportées dans la Table 1.

| Amorces à obstruante [+voisé] | | | | | |
|---|---|---|---|---|---|
| Appariement de voisement congruent | | | Appariement de voisement incongruent | | |
| *contrôle* | en paire minimale | CIBLE | *contrôle* | en paire minimale | CIBLE |
| *grossi* | briguer | mandat | *grossi* | briguer | nettoyer |
| *jumelle* | désert | sable | *jumelle* | désert | chocolat |
| *bandits* | revue | magazine | *bandits* | revue | non |
| *obus* | abat | viande | *obus* | abat | poisson |
| *pincer* | gaver | oie | *pincer* | gaver | erreur |
| *lingot* | cajou | noix | *lingot* | cajou | bonbon |
| *entraver* | aggraver | pire | *entraver* | aggraver | feuilles |
| *douleur* | modèle | mannequin | *douleur* | modèle | hôtel |
| *outils* | égaux | pareils | *outils* | égaux | montagne |
| *héros* | envie | désir | *héros* | envie | cours |
| *guidon* | badaud | passant | *guidon* | badaud | mer |
| *talent* | radeau | méduse | *talent* | radeau | pelle |
| *fumer* | caser | ranger | *fumer* | caser | briser |
| *éduquer* | embauché | travail | *éduquer* | embauché | argent |
| *abouti* | avaler | déglutir | *abouti* | avaler | canapé |
| *tampon* | combat | boxe | *tampon* | combat | math |
| *quittance* | pédale | vélo | *quittance* | pédale | fleur |
| *impliquer* | embraser | feu | *impliquer* | embraser | bisou |
| *pinceaux* | cabot | chien | *pinceaux* | cabot | voiture |
| *verrue* | condé | flic | *verrue* | condé | fromage |

| Amorces à obstruante [–voisé] | | | | | |
|---|---|---|---|---|---|
| Appariement de voisement congruent | | | Appariement de voisement incongruent | | |
| *contrôle* | en paire minimale | CIBLE | *contrôle* | en paire minimale | CIBLE |
| *grossi* | briquer | nettoyer | *grossi* | briquer | mandat |
| *jumelle* | dessert | chocolat | *jumelle* | dessert | sable |
| *bandits* | refus | non | *bandits* | refus | magazine |
| *obus* | appat | poisson | *obus* | appat | viande |
| *pincer* | gaffer | erreur | *pincer* | gaffer | oie |
| *lingot* | cachou | bonbon | *lingot* | cachou | noix |
| *entraver* | agrafer | feuilles | *entraver* | agrafer | pire |
| *douleur* | motel | hôtel | *douleur* | motel | mannequin |
| *outils* | echo | montagne | *outils* | echo | pareils |
| *héros* | amphi | cours | *héros* | amphi | désir |
| *guidon* | bateau | mer | *guidon* | bateau | passant |
| *talent* | rateau | pelle | *talent* | rateau | méduse |
| *fumer* | casser | briser | *fumer* | casser | ranger |
| *éduquer* | empocher | argent | *éduquer* | empocher | travail |
| *abouti* | affalé | canapé | *abouti* | affalé | déglutir |
| *tampon* | compas | math | *tampon* | compas | boxe |
| *quittance* | pétale | fleur | *quittance* | pétale | vélo |
| *impliquer* | embrasser | bisou | *impliquer* | embrasser | feu |
| *pinceaux* | capot | voiture | *pinceaux* | capot | chien |
| *verrue* | comté | fromage | *verrue* | comté | flic |

TABLE 1 : Listes des stimuli amorce–cible et contrôle–cible utilisés dans les tests

30 couples amorce–cible constituées de mots non reliés sémantiquement (par ex. nuage–PION) ont été ajoutés aux stimuli test. La tâche des sujets étant de décider si la séquence de lettre (la cible) affichée à l'écran (et non le mot amorce entendu) était un mot du lexique français ou non, ont été créés 50 couples de mot amorce–pseudo-mot cible, dont 10 dérivés d'un mot réel relié sémantiquement à l'amorce (par ex. tomate–ROUZE) et 40 sans lien sémantique potentiel (par ex. visage–PLAME). Ces pseudo-mots ont été créés sur la base d'un mot réel présentant une substitution ou une permutation de lettres. A terme, les couples de mots composés d'une amorce en paire minimale de voisement associée à une cible reliée sémantiquement représentaient 20 % des stimuli de chaque liste proposée à chaque sujet. Afin que les sujets ne soient jamais soumis deux fois à la même amorce ou à la même cible, les stimuli ont été répartis en 8 listes affectées chacune à 8 groupes de 17 sujets différents (N = 136). L'expérience durait environ 30 mn.

L'expérience en parole chuchotée de la première étude (Meynadier & Dufour 2016) est reconduite ici.[1] Comme elle, les nouveaux tests portent sur les 8 conditions présentées dans la Table 1 : (i) deux où l'amorce en paire minimale est reliée sémantiquement à la cible congruente en appariement de voisement (pé*d*ale–VÉLO, pé*t*ale–FLEUR) ; (ii) deux où l'amorce en paire minimale est sans lien sémantique avec la cible qui correspond à celle reliée sémantiquement de l'autre membre de la paire minimale de voisement (appariement de voisement incongruent : pé*d*ale–FLEUR, pé*t*ale–VÉLO) ; (iii) quatre amorces contrôles pour chaque condition (i) et (ii) où aucun lien ne relie l'amorce et la cible (quittance–VÉLO, quittance–FLEUR). Seule la durée du délai entre l'amorce audio et la cible visuelle allongée à 50 ms (au lieu de 0 ms) diffère de la première étude. La taille de ce délai inter-stimuli est la plus fréquemment employée dans les études utilisant le paradigme de l'amorçage phonologique (par ex., où une amorce partage des phonèmes avec sa cible, cf. Dufour 2008 pour une revue).

## 2.2 Participants et passation des tests

Les participants, différents mais comparables à ceux de la première étude, ont été rémunérés. La passation des tests a suivi exactement les mêmes conditions que celle de l'étude de 2016, excepté

[1] L'expérience en parole modale (4 listes pour 4 groupes x 17 sujets, N = 68), où seules les deux conditions avec appariement de voisement congruent et les deux contrôles ont été testées, ne sont pas détaillées ici. Les résultats de ces tests sont rapportés dans la discussion générale.

que le délai entre l'amorce audio et la cible visuelle a été fixé à 50 ms.

## 2.3 Mesures et analyses statistiques

Les résultats ont été analysés au moyen d'un modèle linéaire généralisé à effets mixtes (Baayen 2008). Les analyses ont porté sur les temps de réaction (RT en ms) incluant uniquement les réponses correctes. Par ailleurs, les RT supérieurs à 1300 ms ont été exclus. Afin que le modèle respecte les critères de normalité de la distribution des résidus ainsi que l'homogénéité de la variance, une transformation logarithmique a été appliquée sur les RT (Baayen & Milin 2010), puis pour chaque participant les RT supérieurs et inférieurs à 2,5 écart-type de leur moyenne ont été également écartés. 4,52% des données ont été ainsi rejetées. Les valeurs rapportées pour F et *p* ont été obtenues à l'aide du package *lmertest* du logiciel R (Kuznetsova et al. 2016).

Le modèle incluait le *Voisement* ([+voisé], [–voisé]), l'*Appariement de voisement* (congruent, incongruent) et le *Type d'amorce* (en paire minimale, contrôle) comme facteurs à effets fixes (Table 1). La structure aléatoire du modèle incluait un intercept différent pour les participants et les items, ainsi que des pentes différentes à la fois par participants et par items pour le facteur *Type d'amorce*. Les comparaisons deux à deux, permettant d'évaluer les effets d'amorçage, ont été réalisées à l'aide de la fonction *glht* du package *multcomp* (Bretz et al. 2011) avec une correction de Bonferroni. Les RT moyens ainsi que les pourcentages de réponses correctes obtenus dans chaque condition sont présentés dans la Figure 1.

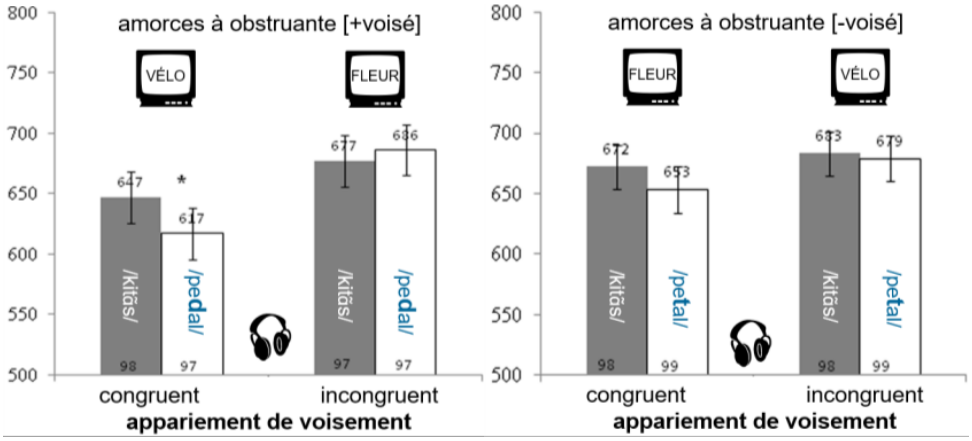



Figure 1 : RT moyen (ms) et erreur-type des réponses pour les amorces chuchotées contrôles (barres grises) vs en paire minimale à obstruantes [+voisé] et [–voisé] (barres blanches) selon les conditions d'appariement de voisement (congruent vs incongruent). Les pourcentages de réponses correctes sont donnés au bas des barres. Un astérisque (*) indique un effet d'amorçage significatif.

## 2.4 Résultats

Le modèle révèle un effet simple du *Type d'amorce* (paire minimale, contrôle) [F=5.22, p<.05] avec des RT plus rapides sur les mots cibles précédés d'une amorce en paire minimale que d'une amorce contrôle. Les effets simples de *Voisement* ([+voisé], [–voisé]) et d'*Appariement de voisement* (congruent, incongruent) n'atteignent pas la significativité. Seule l'interaction *Type d'amorce** *Appariement de voisement* est significative [F=6.64, p<.05]. De façon globale, cette interaction est due à un effet d'amorçage plus important dans le cas d'un appariement congruent que dans le cas d'un appariement incongruent. La double interaction *Voisement*Type d'amorce** *Appariement de voisement* n'est pas significative [F=1.86, p=.18].

Cependant, de façon cruciale, les sous-analyses dans chaque condition de *Voisement* ([+voisé] et

[–voisé]) révèlent une interaction significative *Type d'amorce*Appariement de voisement* seulement pour les amorces à obstruante [+voisé] (pé*d*ale) [F=6.56, p<.05] et non pour celles à obstruante [–voisé] (pé*t*ale) [F=0.44, p>.20]. Pour les amorces à obstruante [+voisé], les comparaisons deux à deux avec une correction de Bonferroni montrent un effet d'amorçage significatif seulement pour les couples amorce–cible appariés en voisement (congruent, pé*d*ale–VÉLO) [z=3.03, p<.05]. Dans ce cas, le RT (617 ms) est en moyenne 30 ms plus rapide que dans la condition contrôle (647 ms, quittance–VÉLO). En revanche, aucun effet d'amorçage n'a été observé pour les couples amorce–cible non appariés en voisement (incongruent, pé*d*ale–FLEUR comparé au contrôle quittance–FLEUR) [z=-1.01, p>.20]. Pour les amorces à obstruante [–voisé], aucun effet d'amorçage n'est observé, tant pour les couples amorce–cible appariés en voisement (congruent, pé*t*ale–FLEUR comparé au contrôle quittance–FLEUR) [z=2.06, p>.20] que pour ceux non appariés (incongruent, pé*t*ale–VÉLO comparé au contrôle quittance–VÉLO) [z=0.92, p>.20].

Pour les pourcentages de réponses correctes, un modèle *logit* à effets mixtes n'a montré aucun effet.

En résumé, contrairement aux résultats de notre première étude où amorce et cible se suivaient sans délai, un délai de 50 ms permet l'émergence d'un effet d'amorçage avec des amorces chuchotées à obstruante [+voisé]. Cependant, l'effet d'amorçage précédemment observé (Meynadier & Dufour 2016) pour les mots amorce à obstruante [–voisé] ne l'est plus avec ce délai de traitement plus long.

# 3 Analyse acoustique des obstruantes voisées vs sourdes

L'analyse porte sur les 40 (2x20) mots en paire minimale de voisement utilisés comme amorce audio dans les expériences de perception et produits en phonation modale et chuchotée par l'un des auteurs de l'étude (Figure 2). Ce sont les mêmes stimuli que ceux de la première étude. Nous renvoyons à Meynadier & Dufour (2016) pour les détails relatifs à leur enregistrement et leur segmentation manuelle. Cette liste de 40 mots comprend 4 paires avec /p–b/, 5 avec /t–d/, 2 avec /k–g/, 5 avec /f–v/, 3 avec /s–z/ et 1 avec /ʃ–ʒ/.

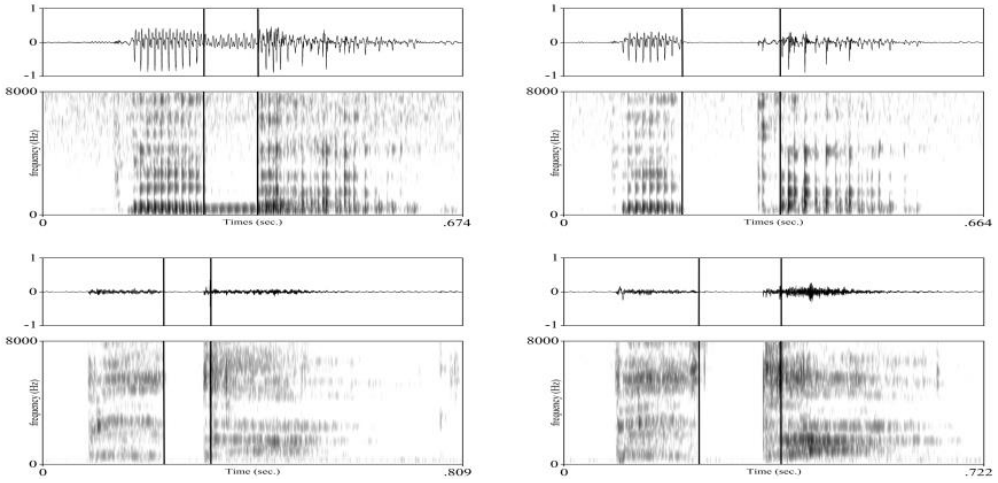



Figure 2 : Onde acoustique et spectrogramme des amorces /pe*d*al/ (à gauche) et /pe*t*al/ (à droite) en parole modale (en haut) et chuchotée (en bas). Les obstruantes /d/ et /t/ sont segmentées par les lignes verticales sombres.

## 3.1 Mesures acoustiques

Les mesures acoustiques effectuées sont les suivantes : (i) Durée de la consonne intervocalique (par ex. /pe*d*al/) ; (ii) Durée de la voyelle préconsonantique (par ex. /p*e*dal/) ; (iii) Intensité RMS de l'explosion ou du bruit de friction de l'obstruante intervocalique ; (iv) Centre de gravité du spectre

de bruit de l'obstruante intervocalique ; (v) Dispersion fréquentielle du spectre de bruit de l'obstruante intervocalique. Ces mesures ont été calculées et extraites sous Praat. La durée des voyelles préconsonantiques a été mesurée du début à la fin du F2, celle des obstruantes de la fin du F2 de la voyelle précédente à la fin de l'explosion ou du bruit de friction. Les mesures spectrales (iii, iv, et v) sont des indices de distribution de l'énergie dans le spectre de bruit de la consonne. Avant extraction, le spectre de bruit a été filtré passe-haut à 1 kHz afin d'exclure les basses fréquences intenses dues au voisement, ne conservant que les composantes du bruit entre 1 et 12,5 kHz. Pour réduire l'influence des voyelles, ces indices spectraux ont été calculés sur les 40 ms au centre des fricatives et sur les 10 ms initiales de l'explosion des plosives. L'intensité RMS (dB) du bruit a été calculée comme l'intensité moyenne sur cet intervalle. Le Centre de gravité (Hz) est le premier moment central (moyenne) du spectre de bruit, c'est-à-dire la fréquence moyenne pondérée en puissance du spectre. La valeur du Centre de gravité augmente avec la quantité d'énergie concentrée dans les hautes fréquences. La Dispersion fréquentielle du spectre de bruit est calculée par la racine carrée du deuxième moment central (variance) du spectre, c'est-à-dire l'écart type de l'énergie répartie dans les fréquences du spectre. Une valeur de dispersion élevée indique une distribution de l'énergie sur une large gamme de fréquences, tandis qu'une valeur faible signale une énergie concentrée autour du Centre de gravité.

### 3.2 Analyse statistique des corrélats acoustiques du voisement

Les 20 amorces à obstruante [+voisé] ont été comparées aux 20 amorces à obstruante [–voisé] en paire minimale par un test de Student sur les cinq paramètres acoustiques présentées ci-avant : durée de l'obstruante, durée de la voyelle préconsonantique, intensité RMS, centre de gravité et dispersion spectrale de l'explosion des plosives ou du bruit de friction des fricatives. Ces analyses portaient sur les productions en parole chuchotée et modale.

**Parole modale** – Les obstruantes voisées sont plus courtes et ont une intensité plus faible (89 ms, 49 dB) que le sourdes (151 ms, 57 dB) [respectivement, $t(38)=7.63$, $p<.0001$ ; $t(38)=2.77$, $p<.01$]. Les voyelles sont plus longues avant une obstruante voisée (154 ms) qu'avant une sourde (118 ms) [$t(38)=3.18$, $p<.01$]. Aucune différence significative ne concerne le Centre de gravité des obstruantes ([+voisé] : 3714 Hz ; [–voisé] : 3960 Hz) [$t(38)=0.59$, $p>.20$], ni la Dispersion fréquentielle ([+voisé] : 1871 Hz ; [–voisé] : 1862 Hz) [$t(38)= 0.07$, $p>.20$].

**Parole chuchotée** – Les obstruantes voisées sont plus courtes (98 ms) que les sourdes (158 ms) [$t(38)=7.06$, $p<.0001$]. Les voyelles sont plus longues avant obstruante voisée (172 ms) qu'avant sourde (134 ms) [$t(38)=4.20$, $p<.001$]. Aucune différence significative n'est observée pour l'intensité RMS ([+voisé] : 52 dB ; [–voisé] : 55 dB) [$t(38)=1.22$, $p>.20$], le Centre de gravité ([+voisé] : 3370 Hz ; [–voisé] : 3896 Hz) [$t(38)=1.16$, $p>.20$] ou la Dispersion fréquentielle du spectre de bruit ([+voisé] : 1416 Hz ; [–voisé] : 1672 Hz) [$t(38)=1.34$, $p=.19$].

En accord avec les études antérieures (cf. §1), pour les stimuli amorces de nos expériences de perception, les corrélats de durée consonantique et vocalique se montrent plus robustes que les corrélats spectraux pour différencier les obstruantes [+voisé] des obstruantes [–voisé] en parole modale comme chuchotée.

## 4 Discussion et conclusion

Par rapport à notre première étude de 2016, l'introduction d'un délai de traitement plus long (50 ms au lieu de 0 ms) entre amorce chuchotée et cible visuelle provoque des effets d'amorçage diffèrents à la fois concernant les mots à obstruante [+voisé] que ceux à obstruante [–voisé].
Concernant les mots chuchotés à obstruante [+voisé], l'étude de 2016 avait mis en évidence que le traitement de la cible sémantiquement reliée présentée immédiatement après la fin acoustique de

l'amorce (pé*d*ale–VÉLO) n'était pas facilité, contrairement à ce qui était observé en parole modale. Cela suggérait que l'auditeur n'avait pu résoudre l'ambiguïté phonétique de l'obstruante [+voisé] produite sans vibration glottique en parole chuchotée. Dans cette nouvelle étude, un effet d'amorçage significatif émerge : la cible reliée sémantiquement à l'amorce est reconnue environ 30 ms plus vite qu'en condition contrôle (quittance–VÉLO). Cela confirme notre hypothèse qu'un temps plus long de traitement est nécessaire à l'auditeur pour identifier l'obstruante [+voisé] chuchotée.

Concernant les mots à obstruante [–voisé], l'étude de 2016 avait mis en évidence que le traitement de la cible sémantiquement reliée présentée immédiatement après la fin acoustique de l'amorce (pé*t*ale–FLEUR) était facilité. Cet effet d'amorçage était d'amplitude similaire (env. 26 ms) à celui observé en parole modale (comparaisons post-hoc par paire avec une correction de Bonferroni ; amorces modales à obstruante [+voisé] : $z=3.49$, $p<.01$ ; à obstruante [–voisé] : $z=3.02$, $p<.05$). Cela suggérait que la propriété sous-jacente [–voisé] était immédiatement disponible pour l'auditeur et que les obstruantes [–voisé] chuchotées n'étaient pas ambiguës. Dans cette nouvelle étude, cet effet d'amorçage n'est plus observé. L'absence d'effet est comparable aux résultats obtenus en parole modale où avec un délai de 50 ms entre amorce et cible les effets d'amorçage disparaissent aussi bien pour les mots à obstruante [–voisé] [$z=2.03$, $p>.20$] que pour ceux à obstruante [+voisé] [$z=1.98$, $p>.05$]. Un tel résultat est très probablement dû au fait que l'effet d'amorçage sémantique est connu pour être éphémère (Marslen-Wilson et al. 1996, Andruski et al. 1994). Ainsi, en augmentant le délai entre amorce et cible, nous prenions le risque de voir cet effet disparaître quand les amorces ne sont pas ambiguës, à savoir pour les amorces modales et pour les amorces chuchotées à obstruantes [–voisé]. Inversement, ce délai de 50 ms entre la fin de l'amorce et le début de la présentation de la cible permet de manière cruciale aux amorces chuchotées 'ambigües', à savoir à obstruante [+voisé], d'activer pleinement leurs représentations lexicales, faisant émerger l'effet d'amorçage.

La résolution tardive de l'ambiguïté des obstruantes chuchotées [+voisé] repose probablement sur un processus d'extraction des indices phonétiques associés au trait de voisement et toujours présents dans le signal chuchoté caractérisé par l'absence de vibration glottique, propriété principale associée au trait [+voisé] en français. L'analyse de ces traces acoustiques (cf. §3), confirmant les études antérieures en parole chuchotée, laisse supposer que les auditeurs peuvent s'appuyer sur des indices résiduels de durée de l'obstruante et/ou de la voyelle qui la précède. Cette étude montre donc que les auditeurs sont pleinement capables d'exploiter les indices résiduels du trait de voisement en parole chuchotée de façon à reconnaitre le mot attendu. De façon cruciale, il est apparu que l'extraction du trait de voisement en parole chuchotée requière un certain temps, probablement nécessaire à la réparation phonologique du trait ou du phonème sous-jacent lors d'une étape pré-lexicale de traitement (Hallé et al. 1998).

# Remerciements

Ce projet de recherche *Whispeech* a été soutenu par les financements ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) et Initiative d'Excellence d'Aix-Marseille Université (A*MIDEX).

# Références

ANDRUSKI J.E., BLUMSTEIN S.E., BURTON M. (1994). The effects of subphonetic differences on lexical access. *Cognition* 52, 163-187.

BAAYEN R.H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

BAAYEN R.H., MILIN P. (2010). Analyzing reaction times. *International Journal of Psychological Research* 3(2), 12-28.

BRETZ F., HOTHORN T., WESTFALL P.H. (2011). *Multiple comparisons using R*. Boca Raton: CRC Press.
DANNENBRING G.L. (1980). Perceptual discrimination of whispered phoneme pairs. *Perceptual and Motor Skills* 51(3), 979-985.
DUFOUR S. (2008). Phonological priming in auditory word recognition: When both controlled and automatic processes are responsible for the effects. *Canadian Journal of Experimental Psychology* 62, 33-41.
FUX T. (2012). *Vers un système indiquant la distance d'un locuteur par transformation de sa voix*. Thèse de doctorat, Université de Grenoble, Grenoble.
GARNIER M., BOUHAKE S., JEANNIN C. (2014). Efforts and coordination in the production of bilabial consonants. *Proceedings of the 10th Int. Seminar on Speech Production*, 138-141. Cologne.
GILICHINSKAYA Y.D. (2012). *Perception of final consonant "voicing" in phonated and whispered speech*. UMI Dissertation Publishing, ProQuest LLC, Ann Arbor, USA.
HALLE P., ANDROJNA K., SEGUI J. (2012). L'assimilation de voisement en français : elle vaut pour les non-mots autant que les mots. Actes des XXIXè JEP, 441-448.
HALLE P., SEGUI J., FRAUENFELDER U., MEUNIER C. (1998). The processing of illegal consonant clusters: a case of perceptual assimilation? *Journal of Experimental Psychology: Human Perception and Performance* 24, 592-608.
ITO T., TAKEDA K., ITAKURA F. (2005). Analysis and recognition of whispered speech. *Speech Communication* 45(2), 139-152.
JOVIČIĆ S.T., ŠARIĆ Z. (2008). Acoustic analysis of consonants in whispered speech. *Journal of Voice* 22, 263-274.
KOHLBERGER M., STRYCHARCZUK P. (2015). Voicing assimilation in whispered speech. *Proceedings of the 18th International Conference on Phonetic Sciences.* Glasgow.
KUZNETSOVA A., BRUUN BROCKHOFF P., HAUBO BOJESEN CHRISTENSEN R. (2016). Tests in linear mixed effect models. Consulté le 20/11/2017, *https://cran.r-project.org/package=lmerTest.*
MALÉCOT A., PEEBLES K. (1965). An optical device for recording glottal adduction-abduction during normal speech. *ZPSK* 18, 545-550.
MARSLEN-WILSON W.D., MOSS H.E., VAN HALEN S. (1996). Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 1376-1392.
MEYNADIER Y. (2015). Aerodynamic tool for phonology of voicing. *Proceedings of the 18th International Conference on Phonetic Sciences*, paper#0497. Glasgow.
MEYNADIER Y., DUFOUR S. (2016). Accès lexical et reconnaissance du voisement en voix chuchotée. XXXIè JEP, 19-27. Paris.
MEYNADIER Y., GAYDINA Y. (2013). Aerodynamic and durational cues of phonological voicing in whisper. *Proceedings of the 14th Interspeech*, 335-339. Lyon.
MILLS T.L.P. (2009). *Speech motor control variables in the production of voicing contrasts and emphatic accent*. Ph Dissertation, University of Edinburg, Ednburgh.
MUNRO M.J. (1990). Perception of "Voicing" in Whispered Stops. *Phonetica* 47(3-4), 173-181.
NETSELL R. (1969). Subglottal and intraoral air pressures during the intervocalic contrast of /t/ and /d/. *Phonetica* 20(24), 68-73.
SLIS L.H. (1970). Articulatory measurements on voiced, voiceless and nasal consonants. *Phonetica* 21(4), 193-210.
TABOSSI P. (1996). Cross-modal semantic priming. *Language and Cognitive Processes* 11, 569-576.
TARTTER V.C. (1989). What's in a whisper? *The Journal of the Acoustical Society of America, 86*, 1678-1683.
VAN DE VELDE D.J., VAN HEUVEN V. (2011). Compensatory strategies for voicing of initial and medial plosives and fricatives in whispered speech in Dutch. *Proceedings of the 17th ICPhS*, 2058–2061. Hong-Kong.
VERCHERAND G. (2010). *Production et perception de la parole chuchotée en français : analyse segmentale et prosodique*. Thèse doctorale, Université Paris VI, Paris.

# Codage efficace à débit variable basé sur la quantification vectorielle à divisions commutées: Application aux paramètres ISF en large bande

Cheraitia Salah Eddine[1], Bouzid Merouane[1] et Meziane Nacéra [2]
(1) Laboratoire Communication Parlée et Traitement du Signal (LCPTS), Université USTHB, BP 32, El-Alia, Bab-Ezzouar, Alger, 16111, Algérie.
(2) Faculté des Hydrocarbures et de la Chimie, Dept. Automatisation et Electrification, Université M'Hamed BOUGARA, Boumerdès, Algérie.
cher.salah@yahoo.fr, mbouzid@usthb.dz, nac.meziane@gmail.com

## RESUME

Le codage efficace des coefficients de prédiction linéaire (LPC) est l'un des problèmes importants dans la conception des codeurs de parole modernes. Les paramètres "fréquences spectrales d'immitance" ISF (Immittance Spectral Frequencies) sont actuellement classés parmi les choix les plus appropriés pour représenter les coefficients LPC en large bande. Dans cet article, nous proposons une version à débit variable du quantificateur vectoriel à divisions commutées (SSVQ) développé pour le codage efficace des paramètres ISF de parole en large-bande, selon des suppositions de transmission à travers un canal non bruité.

## ABSTRACT

**Variable rate switched split vector quantizer for efficient coding of wideband speech ISF parameters.**

Modern speech coders necessitate efficient coding of the linear predictive coding (LPC) coefficients. Immittance Spectral Frequencies (ISF) parameters are currently the most efficient choices of transmission parameters for the LPC coefficients in wideband. In this paper, we propose a variable rate version of the switched split vector quantizer (SSVQ) scheme developed for efficient coding of wideband speech ISF parameters under noiseless channel conditions.

MOTS-CLES : Quantification vectorielle, codage parole en large bande, paramètres ISF, codeur AMR-WB.
KEYWORDS : Vector quantization, wideband speech coding, ISF parameters, AMR-WB coder

# 1 Introduction

Dans la plupart des systèmes de codage de la parole modernes, l'enveloppe spectrale à court terme d'un signal de parole est souvent modélisée par la réponse fréquentielle d'un filtre tout-pôle dont la fonction de transfert est donnée par $H(z) = 1/A(z)$, avec $A(z) = 1 + a_1 z^{-1} + \ldots + a_p z^{-p}$ (Kleijn, Paliwal, 1995). Les coefficients de ce filtre $\{a_i\}_{i=1,2,\ldots p}$, connus aussi sous le nom de coefficients de prédiction linéaire (LPC), sont dérivés du signal d'entrée par une analyse par prédiction linéaire d'ordre $p$ sur chaque trame du signal de parole. La quantification efficace des coefficients LPC joue


10.21437/JEP.2018-16

un rôle important dans la préservation de l'intelligibilité et de la qualité naturelle du signal de parole codé sur toute la largeur de bande. En pratique, ces coefficients ont des propriétés de quantification médiocres. Ainsi, plusieurs représentations paramétriques équivalentes ont été formulées afin de les convertir en paramètres beaucoup plus appropriés à la quantification. Les fréquences de raie spectrales LSF (Line Spectral Frequencies) et les fréquences spectrales d'immitance ISF (Immittance Spectral Frequencies) se sont avérées les plus efficaces pour représenter les coefficients LPC des codeurs de parole modernes basés sur le modèle autorégressif.

Dans ce travail, on s'est intéressé particulièrement à la quantification efficace des paramètres ISF des codeurs de parole en large-bande comme le codeur large-bande à débit multiple adaptatif AMR-WB (Adaptive Multi-Rate Wide-Band) norme ITU-T G.722.2 (Bessette et al., 2002). Notons que les paramètres LSF et ISF sont des représentations équivalentes qui se déduisent mathématiquement les unes des autres par de simples fonctions de conversion. Les paramètres ISF (ISFs), qui ont été introduits par Bistritz et al. (Bistritz et al., 1989), sont définies comme étant les pôles et les zéros d'une fonction d'immitance à la glotte :

$$I_{16}(z) = \frac{A(z) - z^{-16}A(z^{-1})}{A(z) + z^{-16}A(z^{-1})} \quad (1)$$

On obtient alors 16 paramètres ISF et le seizième est un coefficient de réflexion.

Comparé aux codeurs de parole en bande étroite (300–3400 Hz), les codeurs en large-bande ont amélioré le naturel et l'intelligibilité de la parole décodée en élargissant la bande passante utilisée pour la transmission du signal de parole (50–7000 Hz). Cependant, ils requièrent un nombre plus élevé de paramètres LPC, typiquement 16, pour représenter l'enveloppe spectrale de parole. Ainsi, le quantificateur vectoriel (VQ) conventionnel doit fonctionner à des débits plus élevés et sur des vecteurs de dimensions plus grandes si on veut l'utiliser pour le codage des ISFs. Ces exigences pratiques vont conduire à une augmentation excessive de la complexité des calculs et de la taille mémoire. Dans le passé, divers schémas de quantificateurs vectoriels (VQs) structurés ont été développés pour le codage des paramètres ISF en large-bande. Ces schémas peuvent réduire considérablement la complexité des calculs avec une perte modérée des performances de quantification. L'un des schémas de codage les plus présentés dans la littérature est sans doute le quantificateur vectoriel divisé SVQ (Split Vector Quantizer), qui a été d'abord développé par Paliwal et Atal (Paliwal, Atal, 1993) pour les codeurs de parole en bande étroite et exploré par la suite dans le codage de la parole en large-bande (Chen, Wang, 1996). Dans (Biundo et al., 2002) un VQ multi-étages divisé (S-MSVQ) avec prédicteur MA de 1$^{er}$ ordre a été utilisé pour le codage des ISFs en large bande. Le même schéma a été utilisé pour coder les paramètres ISF du AMR-WB (Bessette et al., 2002). Dans (So, Paliwal, 2004), So and Paliwal ont proposé le quantificateur vectoriel à divisions commutées SSVQ (Switched Split Vector Quantizer) et le quantificateur multi-trame à base de GMM (Gaussian mixture model) (So, Paliwal, 2007). Dans (Xiaochen et al., 2009), Xiaochen et al. ont proposé un algorithme efficace de quantification des paramètres ISF basé sur le modèle GMM où les ISFs sont quantifiés par un VQ réseau de point Gaussien. Dans (Sheikhan, Garoucy, 2010), trois schémas hybrides pour la quantification des paramètres ISF en large bande ont été proposés. Le premier schéma est basé sur le SSVQ et le S-MSVQ; le second est basé sur le réseau de neurone SOM (Self Organizing Map) et le troisième est basé sur le S-MSVQ et le réseau de neurone GHSOM (Growing Hierarchical SOM).

Dans ce papier, nous proposons une version à débit variable du quantificateur SSVQ développé pour le codage efficace des paramètres ISF du codeur de parole AMR-WB (G.722.2), selon des

suppositions de transmission à travers un canal non bruité. Les résultats de simulation montreront que notre schéma de codage des ISFs du G.722.2, nommé codeur ISF-VR-SSVQ, peut fournir des performances comparables à celles du SSVQ conventionnel tout en diminuant le débit de 1 à 2 bits par trame.

# 2 Quantificateur vectoriel à divisions commutées

Le quantificateur vectoriel à divisions commutées SSVQ (Switched Split Vector Quantizer) est un schéma de codage hybride conçu à base d'un VQ à commutation combiné avec plusieurs quantificateurs vectoriels divisés (So, Paliwal, 2004), (So, Paliwal, 2007). Tout d'abord, rappelons brièvement les principes de base du VQ conventionnel et du SVQ.

Un VQ, de dimension $k$ et de débit de $R$ bits/échantillon (bpe) est défini comme une fonction d'un espace Euclidien $\Re^k$ vers un dictionnaire fini $Y = \{y_0, \ldots, y_{L-1}\}$ composé de $L = 2^{kR}$ vecteurs-code (Gersho, Gray, 1992). Le principe de conception d'un VQ consiste à décomposer l'espace des vecteurs de source $x$ de dimension $k$ en $L$ classes disjointes $\{R_0,.., R_{L-1}\}$ et associer à chaque classe $R_i$ un vecteur-code unique $y_i$ telle que la distorsion totale moyenne $D$ soit minimisée (Gersho, Gray, 1992). Dans le passé, plusieurs méthodes de conception d'un VQ optimal ont été développées. La plus populaire est sans doute l'algorithme LBG (Linde et al., 1980). Cet algorithme (noté ici LBG-VQ) est une application itérative des deux conditions d'optimalité telle que la partition et le dictionnaire soient mis à jour itérativement. Il converge vers une solution localement optimale selon le choix du dictionnaire initial. Dans nos conceptions à base de systèmes VQ, nous avons utilisé la méthode d'initialisation de Katsavounidis (Katsavounidis et al., 1994) qui nous a permit d'obtenir les meilleurs optimums locaux.

D'autre part, un SVQ de dimension $k$ et de $N$ parties (noté $N$-SVQ) est composé de $N$ quantificateurs VQ classiques de tailles et de dimensions plus petites (Paliwal, Atal, 1993). Son principe de base consiste à partitionner l'ensemble des vecteurs $x$ de dimension $k$ de la base d'apprentissage en $N$ sous-ensembles composés de sous-vecteurs de dimension $k_i$ plus petites (avec $\sum_{i=1}^{N} k_i = k$ ). Ensuite, pour chaque partie, le dictionnaire VQ correspondant sera conçu en utilisant le LBG-VQ. Un $N$-SVQ est donc constitué de $N$ dictionnaires VQ de tailles plus petites $L_i = 2^{Riki}$ (où $L = \prod_{i=1}^{N} L_i$ et $R_i$ est le débit partiel en bpe).

## 2.1 Principe de conception du SSVQ

Le principe de base du SSVQ consiste à diviser l'espace des vecteurs de la base d'apprentissage en plusieurs parties disjointes (régions de commutation), où chaque partie est représentée par un SVQ local approprié. La figure 1 présente le schéma bloc du principe de construction du dictionnaire SSVQ.

La première étape consiste à appliquer l'algorithme LBG-VQ sur toute la base d'apprentissage afin de produire $m$ vecteurs-code. L'ensemble de ces vecteurs-code est appelé dictionnaire VQ commutateur $Y_m$ où $m$ représente le nombre de direction de commutation. Ensuite, ce dictionnaire sera utilisé pour partager la base d'apprentissage en $m$ parties suivant le critère du voisin le plus proche. Dans la deuxième étape, chaque partie $i$ ($i$ = 1,…, $m$) sera représentée par un $N$-SVQ$i$ local correspondant.

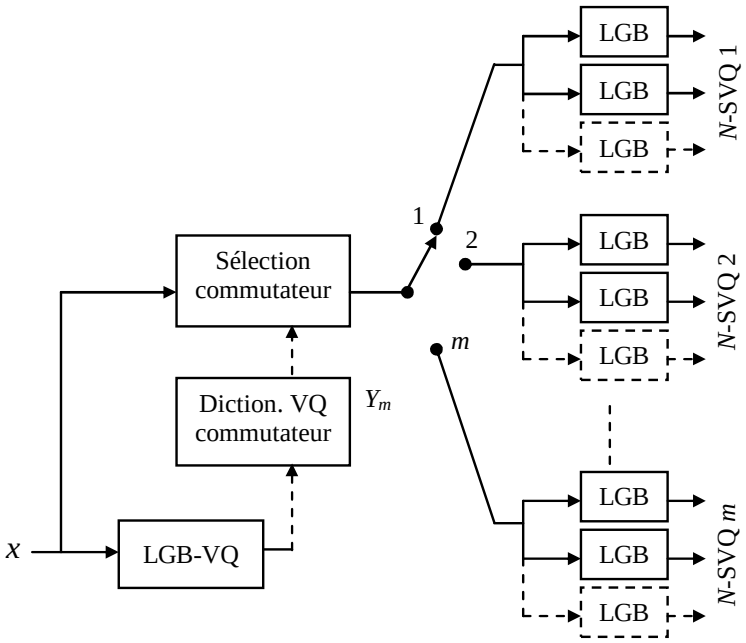



Figure 1 : Construction du dictionnaire SSVQ

### 2.2 Codage/décodage SSVQ

Le codage SSVQ d'un vecteur de source $x$ passe par deux étapes. Au début, le vecteur $x$ est commuté vers l'une des $m$ directions possibles. Par la suite, ce vecteur sera quantifié par le $N$-SVQ local correspondant qui a été sélectionné par $Y_m$. Ainsi, le codeur SSVQ transmet au décodeur un indice $i$ composé de $N + 1$ indices binaires concaténés. Le premier indice $i_s$ ($s = 1,\ldots, m$) indique la direction de commutation et les $N$ indices binaires $i_n$ ($n = 1,\ldots, N$) restants sont fournit par le $N$-SVQ$i_s$ correspondant à la direction $i_s$. Pour un SSVQ de $b$ bits/vecteur et $m$ directions de commutation, on requière $b_s = \log_2(m)$ bits pour identifier séparément toute les directions de commutation possibles. Le débit restant de $(b - b_s)$ bits sera partagé en $N$ débits partiels pour coder les $N$ sous-vecteurs de $x$ par le $N$-SVQ correspondant. Ces débits partiels sont notés par $b_j$ ($j = 1\ldots N$).

Le décodeur SSVQ, qui possède les mêmes dictionnaires que ceux du codeur, reçoit l'indice envoyé $i = (i_s\ i_n)$ avec $n = 1,\ldots, N$. Il utilise le premier indice $i_s$ pour sélectionner la direction de commutation. Ensuite, il construit le vecteur décodé de $x$ en concaténant les sous vecteurs-code d'indices $i_n$ correspondants au $N$-SVQ de la partie $i_s$.

# 3 Codage efficace des paramètres ISF du codeur AMR-WB

Dans cette section, nous proposons un schéma de codage efficace des paramètres ISF du standard AMR-WB (Rec. G.722.2) pour des transmissions à travers un canal non bruité. Il s'agit d'une version à débit variable du SSVQ conventionnel développée en exploitant la stabilité de l'indice de la direction de commutation pour plusieurs trames successives. Mais d'abords, nous présentons les performances du schéma SSVQ conventionnel appliqué au codage des ISFs du G.722.2.

## 3.1 Codage des paramètres ISF par SSVQ

Nous présentons, ci-dessous, les performances du codeur des ISFs du G.722.2, conçu à base de la technique SSVQ et appelé "codeur ISF-SSVQ". Les performances de quantification de nos codeurs des ISFs sont évaluées par la distorsion spectrale SD (Spectral Distorsion). L'expression de la SD pour une trame *i* est donnée en décibels par (Paliwal, Atal, 1993), (So, Paliwal, 2007) et (Cheraitia, Bouzid, 2014):

$$SD_i = \sqrt{\frac{1}{n_1 - n_0} \sum_{n=n_0}^{n_1-1} \left[ 10 \log_{10} \frac{S(e^{j2\pi n/N})}{\hat{S}(e^{j2\pi n/N})} \right]^2} \quad , \tag{2}$$

où $S(e^{j2\pi n/N})$ et $\hat{S}(e^{j2\pi n/N})$ représentent respectivement les spectres de puissance original et quantifié du filtre de synthèse LPC de la $i^{\text{ème}}$ trame du signal de parole.

En général, une quantification de qualité transparente est obtenue si les trois conditions suivantes sont maintenues (Paliwal, Atal, 1993) : **1)-** la distorsion spectrale (SD) moyenne est d'environ 1 dB, **2)-** aucune trame externes "outliers" ne doit avoir une SD qui dépasse les 4 dB et **3)-** le pourcentage des trames outliers ayant une SD entre 2 et 4 dB est moins de 2%. Selon Guibé et al. (Guibé et al., 2001) et Cheraitia et Bouzid (Cheraitia, Bouzid, 2014), les tests d'écoute ont montré que ces conditions de quantification transparente, qui sont souvent utilisés dans le cas du codage de la parole en bande étroite, sont aussi valables dans le cas du codage en large-bande.

La base de données parole utilisée dans ce travail se compose d'environ 85 minutes de parole prise de la base de données internationale TIMIT, avec une fréquence d'échantillonnage de 16 kHz (DARPA, 1988). Pour construire la base des vecteurs ISF, nous avons utilisé la même fonction d'analyse LPC du G.722.2 (Bessette et al., 2002) où une analyse LPC d'ordre 16, par la méthode d'autocorrélation, est effectuée sur chaque trame d'analyse de 20 ms. Une partie de la base de données ISF (208363 vecteurs ISF) est utilisée pour l'apprentissage et la partie restante, de 48606 vecteurs ISF (différente de la base d'apprentissage), est utilisée pour les tests.

Afin d'améliorer davantage les performances de nos codeurs des ISFs et obtenir une quantification transparente à des débits plus bas, nous avons utilisé une mesure de distance euclidienne pondérée plus appropriée (Paliwal, Atal, 1993) et (Cheraitia, Bouzid, 2014) :

$$d(f, \hat{f}) = \sum_{i=1}^{16} [w_i (f_i - \hat{f}_i)]^2 \, , \tag{3}$$

où $f_i$ et $\hat{f}_i$ sont respectivement les $i^{\text{ème}}$ coefficients des ISFs original $f$ et quantifié $\hat{f}$ et $w_i$ représente le poids spectral (Paliwal, Atal, 1993) assigné au $i^{\text{ème}}$ coefficient du vecteur ISF.

Pour différents débits de codage $b$ ($b_s$+ $b_1$ +…+ $b_5$), les performances d'un exemple de codeur ISF-SSVQ de $m$ = 16 directions de commutation sont données dans la Table 1. Notons que dans la conception de notre codeur ISF-SSVQ, les vecteurs ISF large-bande de dimension 16 sont divisés en 5 parties suivant la division (3 – 3 – 3 – 3 – 4) et, dans la mesure du possible, les bits sont uniformément alloués dans chaque régions.

| Débit - Bits/trame | SD Moy. | "Outliers" (en %) | |
|---|---|---|---|
| $b$ ($b_s$+$b_1$ +…+ $b_5$) | (dB) | 2-4 dB | > 4 dB |
| **46** (4+9+9+8+8+8) | 0.98 | 0.71 | 0.00 |
| **45** (4+9+8+8+8+8) | 1.01 | 0.86 | 0.00 |
| **44** (4+8+8+8+8+8) | 1.03 | 0.91 | 0.00 |
| **43 (4+8+8+8+8+7)** | **1.08** | **1.63** | **0.00** |
| **42** (4+8+8+8+7+7) | 1.12 | 2.12 | 0.00 |
| **41** (4+8+8+7+7+7) | 1.16 | 2.72 | 0.00 |

TABLE 1 : Performances des codeurs ISF-SSVQ de 5 parties ($m$ = 16)

Ces résultats de simulation montrent que le codeur ISF-SSVQ de 5 parties, avec la distance pondérée, peut réaliser une quantification de qualité transparente à 43 bits/trame.

## 3.2 SSVQ à débit variable pour le codage des ISFs en large bande

Dans cette section, nous proposons une version à débit variable du SSVQ appliqué au codage efficace des paramètres ISF du G.722.2, selon des suppositions de transmission idéales à travers un canal non bruité.

Le nouveau schéma, nommé ISF-VR-SSVQ, garde pratiquement le même concept de construction du SSVQ de base. Le changement se fait au niveau du codage/décodage où l'on exploite la stabilité de l'indice de commutation $i_s$ pour plusieurs trames successives. En effet, une étude statistique de la variation de cet indice (c.f., Figure 2) a montrée que plusieurs vecteurs ISF de trames successives peuvent être codés avec le même quantificateur local $N$-SVQ de direction $i_s$. Cette observation nous a donnée l'idée de concevoir un SSVQ à débit variable en ajoutant un bit supplémentaire à l'ensemble des bits transmis afin d'indiquer au décodeur de maintenir l'indice $i_s$ précédent (sans le transmettre) ou d'utiliser un autre indice qui sera transmis.

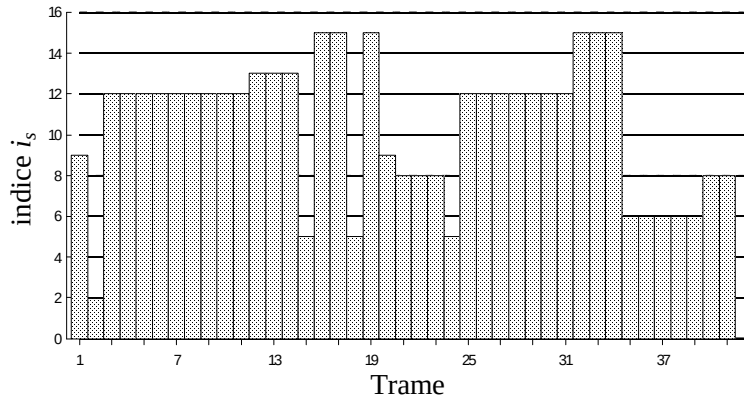



Figure 2: Exemple de variation de l'indice de commutation durant le codage SSVQ

Ainsi, si la direction de commutation actuelle est identique à celle de la trame précédente, le codeur SSVQ transmet seulement un bit supplémentaire "0" sans transmettre l'indice $i_s$. Dans ce cas, une diminution du débit de $b_s - 1$ bits/trame peut être obtenue lors de la transmission comparé au codeur ISF-SSVQ conventionnel. Dans le cas contraire, le codeur transmet un bit supplémentaire "1" en plus du nouveau indice $i_s$; ce qui laisse penser que cette idée peut causer une augmentation du débit de codage. Ceci n'étant pas vrai puisque les résultats de simulation, présentés ci-dessous, montreront que le codeur ISF-VR-SSVQ peut assurer des performances comparables à celles du ISF-SSVQ avec une réduction du débit de codage.

Les performances de notre codeur ISF-VR-SSVQ de 5 parties sont présentées dans la Table 2. Notons que les codeurs ISF-VR-SSVQ de 5 parties ont été conçus dans les mêmes conditions que les codeurs ISF-SSVQ de 5 parties (c-à-d., $m$ = 16, division, allocation de bits, base de données ISF, distance pondérée). Cependant, le débit de codage réel est déterminé à la fin de la transmission car il dépend de la stabilité de la valeur de l'indice de commutation d'une trame vers une autre, comme expliqué précédemment.

| Débit Bits/trame | SD Moy. (dB) | "Outliers" (en %) | |
|---|---|---|---|
| | | 2-4 dB | > 4 dB |
| **44.79** | 0.98 | 0.71 | 0.00 |
| **43.79** | 1.01 | 0.86 | 0.00 |
| **42.79** | 1.03 | 0.91 | 0.00 |
| **41.79** | 1.08 | 1.63 | 0.00 |

TABLE 2 : Performances des codeurs ISF-VR-SSVQ de 5 parties ($m$ = 16)

En comparant ces résultats avec ceux données dans la Tables 1, on remarque clairement que le codeur ISF-VR-SSVQ assure pratiquement les mêmes performances que ceux du codeur ISF-SSVQ mais avec une diminution du débit de codage. En effet, le codeur ISF-VR-SSVQ de 5 parties a besoin de seulement de 41.79 bits/trame pour réaliser une quantification de qualité transparente.

# 4 Conclusion

Dans ce travail, une version à débit variable du SSVQ a été développée pour le codage efficace des paramètres ISF en large bande du codeur G.722.2, selon des suppositions de transmission à travers un canal non bruité. Les résultats de simulation ont montré que notre codeur ISF-VR-SSVQ fournit des performances comparables à celles du ISF-SSVQ conventionnel tout en assurant un gain de 1-2 bits/trame. Les performances du codeur ISF-VR-SSVQ en présence des erreurs de canal reste à être étudiées.

# Références

BESSETTE B., SALAMI R., LEFEVRE R., JELINEK M., ROTOLA-PUKKILA J., VAINIO J., MIKKOLA H., JARVINEN K. (2002). The adaptive multirate wideband speech codec (AMR-WB). *IEEE Transactions on Speech and Audio Processing*, Vol.10, no. 8. 620-636.

BISTRITZ Y., LEV-ARI H., KAILATH T. (1989). Immittance domain levinson algorithms. *IEEE Transactions on Information Theory,* Vol. 35, 675–682.

BIUNDO G., GRASSI S., ANSORGE M., PELLANDINI F., FARINE P. A. (2002). Design techniques for spectral quantization in wideband speech coding. *Proceedings of 3rd COST 276 Workshop on Information and Knowledge Management for Integrated Media Communication, Budapest,* 114-119.

CHEN J. H., WANG D. (1996). Transform predictive coding of wideband speech signals. *In Proc. of the ICASSP'96,* Atlanta, USA, 275–278.

CHERAITIA S., BOUZID, M. (2014). Robust coding of wideband speech immittance spectral frequencies. *Speech Communication, Elsevier,* Vol. 65. 94-108.

*DARPA TIMIT Acoustic-phonetic Continuous Speech Database*, Technology Building, National Institute of Standards and Technology (NIST), Gaithersburg October 1988.

GERSHO A., GRAY R. M. (1992). *Vector quantization and Signal compression*, Kluwer Academic Publishers, USA.

GUIBÉ G., HOW H. T., HANZO L. (2001). Speech spectral quantizers for wideband speech coding. *European Transactions on Telecommunications,* Vol. 12, no. 6, 535-545.

KATSAVOUNIDIS I., KUO C., ZHANG Z. (1994). A new initialization technique for generalized Lloyd iteration. *IEEE Signal Proc. Letters,* Vol. 1, 144-146.

KLEIJN W. B., PALIWAL K. K. (1995). *Speech coding and synthesis*. Elsevier Science B.V.

LINDE Y., BUZO A., GRAY R. M. (1980). An Algorithm for Vector Quantization Design, *IEEE Transactions on Communications,* Vol. 28, 84-95.

PALIWAL K. K., ATAL B. S. (1993). Efficient vector quantization of LPC parameters at 24 bits/frame. *IEEE Transactions on Speech and Audio Processing*, Vol. 1, no. 1, 3-14.

SHEIKHAN M., GAROUCY S. (2010). Hybrid VQ and neural models for ISF quantization in wideband speech coding. *World Applied Sciences Journal,* Vol. 10, 59-66.

SO S., PALIWAL K. K. (2004). Efficient vector quantization of line spectral frequencies using the switched split vector quantiser. *Proceedings of Int. Conf. Spoken language Processing,* Jeju, Korea.

SO S., PALIWAL K. K. (2007). Comparative study of LPC parameter representations and quantisation schemes for wideband speech coding. *Digital Signal Processing,* Vol. 17, 114-137.

XIAOCHEN W., YONG Z., RUIMIN H., XI D. (2009). An Immittance Spectral Frequency parameters quantization Algorithm based on Gaussian Mixture Model. In *Proceedings of International Conference on Multimedia Information Networking and Security (MINES'09),* 324-328.

# Développement de la parole et de la mastication : Evolution de la durée des cycles oscillatoires mandibulaires observés entre 8 et 14 mois chez 4 enfants québécois.

Leslie Lemarchand[1, 2], Andrea A.N MacLeod [2], Mélanie Canault [1], Sophie Kern [1]
(1) Laboratoire Dynamique Du Langage (CNRS/Université Lyon 2-UMR 5596),
14 avenue Berthelot ; 69363 Lyon Cedex 07, France
(2) Ecole d'orthophonie et d'audiologie de l'Université de Montréal,
7077 avenue du Parc, QC H3N 1X7 Montréal, Canada
leslie.lemarchand@univ-lyon2.fr, andrea.macleod@umontreal.ca,
melanie.canault@univ-lyon1.fr, sophie.kern@univ-lyon2.fr

## RESUME

La mastication et la parole sont deux activités motrices acquises qui apparaissent au cours de la première année de vie grâce à l'émergence d'oscillations rythmiques mandibulaires. Cette étude vise à investiguer l'évolution des durées des cycles oscillatoires mandibulaires pour ces deux activités entre 8 et 14 mois et d'observer comment ces patrons interagissent au cours du développement. Pour cela, 4 enfants français québécois ont été enregistrés à 8, 10, 12 et 14 mois au cours d'activités de parole spontanée et de nutrition. Les analyses acoustiques et vidéos effectuées mettent en évidence une différence significative entre la durée de la syllabe et celle d'un cycle masticatoire. Cette distinction temporelle pourrait illustrer la spécialisation précoce de l'organisation temporelle des activités de parole et de nutrition.

## ABSTRACT

**Speech and feeding development: A longitudinal study on 4 Quebecois French-speaking children between 8 and 14 months.**

Speech production and chewing are two motor activities characterized by rhythmic jaw oscillation, which appear throughout the first year of life. The aim of this study was to investigate the evolution of jaw temporal oscillation patterns during nutrition and speech between 8 and 14 months and to determine in which ways these patterns could interact. To address this issue, we monthly recorded 4 Canadian-French-speaking children during speech and meal activities at 8, 10, 12 and 14 months of age. Acoustic and video analyses indicate a significant difference between syllable duration and chewing cycle duration. These results could reflect the early mandible temporal rhythmic organization for chewing and speech.

MOTS-CLES : babillage, mastication, acquisition, oscillations mandibulaires.
KEYWORDS: babbling, chewing, acquisition, jaw temporal oscillation patterns.


10.21437/JEP.2018-17

# 1 Introduction

La mastication et la parole sont des activités qui s'acquièrent progressivement au cours de la petite enfance grâce à l'interaction de différents facteurs intrinsèques (e.g. développement moteur, contraintes physiques et physiologiques) et extrinsèques (i.e. environnement) à l'enfant. D'un point de vue moteur, ces activités se caractérisent par une alternance rythmique et continue de cycles d'ouverture et de fermeture de la mandibule (MacNeilage, 1998). L'émergence de ce geste mandibulaire serait à l'origine de la transition entre l'alimentation exclusivement lactée (i.e oralité alimentaire primaire) et l'alimentation diversifiée (i.e. oralité alimentaire secondaire) (Le Révérend et al., 2014) ainsi que de la production des premières syllabes marquant l'entrée dans le babillage (MacNeilage, 1998).

## 1.1 L'accélération du rythme oscillatoire mandibulaire : un indice de développement des compétences oro-motrices

Bien que peu nombreuses, les données de la littérature décrivent une accélération du rythme oscillatoire mandibulaire en fonction de l'âge pour les activités de nutrition et de parole. En effet, pour la nutrition, des études vidéos et cinématiques rapportent une augmentation de l'efficacité masticatoire entre 6 et 24 mois (Gisel, 1991) en partie expliquée par l'accélération de la fréquence oscillatoire mandibulaire entre la petite enfance (0,8-1,2 Hz (Goldfield & Wolff, 2003 ; Le Révérend et al., 2014)) et l'âge adulte (1,5- 3 Hz (Jürgens, 1998)). Pour la parole, cette fréquence oscillatoire passe de 3 Hz (Bickley et al., 1986) vers 7 mois, à 5-6 Hz à l'âge adulte (Jürgens, 1998). La syllabe pouvant être considérée comme le résultat d'un cycle oscillatoire mandibulaire (MacNeilage, 1998), le timing des déplacements mandibulaires peut également être inféré à partir de l'observation de la durée de la syllabe. Ainsi, une diminution de la durée syllabique au cours du développement a été mise en évidence à plusieurs reprises (Kent & Murray, 1982; Dolata, Davis, & MacNeilage, 2008). Les patrons temporels mandibulaires témoigneraient de l'immaturité précoce du contrôle moteur lors de l'apparition de la mastication et du babillage (Smith & Goffman, 1998) et de l'amélioration des compétences oro-motrices au cours du développement (Smith & Zelaznik, 2004).

## 1.2 Une interaction entre l'évolution des patrons temporels mandibulaires de la mastication et du babillage ?

Outre le rôle prépondérant des oscillations mandibulaires dans l'apparition et le développement des compétences articulatoires et masticatoires, le fait que les activités de parole et de nutrition reposent sur des effecteurs anatomiques et physiologiques communs et que des troubles alimentaires et langagiers coexistent chez certains enfants (Malas et al., 2015) pose l'hypothèse d'une interaction entre le développement de la mastication et de la parole. L'objectif de cette étude est ainsi de comparer de manière longitudinale l'évolution des durées des cycles oscillatoires mandibulaires pour les activités de parole et de mastication chez des enfants âgés de 8 à 14 mois et de voir comment ces patrons interagissent au cours du développement.

# 2 Méthode

La réalisation de cette étude a été approuvée par le comité d'Ethique de la Recherche du CHU Sainte Justine (Montréal, Canada).

## 2.1 Participants

4 enfants français québécois nés à terme (2 filles et 2 garçons) âgés de 8 et 9 mois (plus ou moins 15 jours) ont participé à cette étude. Aucun des sujets ne présentait de troubles moteurs, auditifs ou mentaux, ni d'antécédents médicaux concernant la sphère orale (malformation orale, ventilation et/ou alimentation artificielle, troubles alimentaires).

## 2.2 Procédure

Chaque participant a été enregistré 4 fois, soit une session à 8 (ou 9 mois), 10, 12 et 14 mois. Toutes les sessions d'enregistrement ont été réalisées au sein du centre de Réadaptation Marie-enfant du CHU Sainte- Justine (Montréal, Canada). Chaque session impliquait des enregistrements audio et vidéos réalisés en chambre sourde à l'aide d'un microphone directionnel de haute qualité (Sennheiser K6) et d'une caméra Microsoft LifeCam Studio (30 images par seconde, résolution 640x480). Les sujets étaient placés face à la caméra dans une chaise haute. L'expérimentation, d'une durée moyenne de 45 minutes pour chacune des sessions, a donc permis de recueillir deux types de données :

- Des enregistrements vidéos des mouvements masticatoires

Les participants étaient enregistrés au cours d'un repas, afin de déterminer le nombre des cycles masticatoires au cours d'une bouchée ainsi que leur durée en fonction des textures administrées. Les aliments présentés au cours des repas ont été choisis à partir des textures standardisées issues du test « Schedule for Oral-Motor Assessment » (SOMA) (Reilly et al., 2000) qui évalue les compétences oro-motrices du jeune enfant. Ainsi, les textures « purées », « biscuits », « semi-solides » (e.g. pain de mie) et « solides » (e.g. morceaux de pomme) ont été proposées aux participants. Les parents avaient pour consigne de donner à manger à leur enfant de la manière la plus habituelle possible en utilisant les cuillères mises à disposition par les expérimentateurs afin d'uniformiser la taille des cuillérées.

- Des enregistrements audio des productions babillées

Les productions babillées spontanées des participants au cours d'interactions avec les parents et les expérimentateurs ont été enregistrées pour le recueil des syllabes.

## 2.3 Traitement des données

- Analyse de la durée des cycles masticatoires

Les enregistrements vidéos ont été analysés à l'aide du logiciel Datavyu®. La durée d'un cycle masticatoire (e.g. une ouverture-fermeture de la mandibule) a été calculée en suivant la méthodologie utilisée dans l'étude de Gisel (1991). Cette durée a ainsi été obtenue en effectuant un ratio entre la durée d'une séquence masticatoire et le nombre de cycles masticatoires effectués au cours de cette séquence. La durée d'une séquence masticatoire ayant été définie comme la période située entre la première ouverture mandibulaire après le placement du bol alimentaire et la dernière fermeture mandibulaire précédent la déglutition (Steeve & Moore, 2009). Les données de cette étude regroupent les durées moyennes des cycles masticatoires obtenues pour les textures « semi-solides » et « biscuits » (TABLE 1).

- Analyse des durées syllabiques

L'analyse des données acoustiques a été réalisée avec le logiciel Praat®. Les séquences babillées recueillies ont été extraites puis segmentées en syllabes et annotées en fonction de leur type (monosyllabe, pluri-syllabe) et de leur structure (e.g. consonne-voyelle, voyelle-consonne-voyelle). La durée d'un cycle oscillatoire mandibulaire au cours de la parole a été inférée à partir des durées syllabiques obtenues pour les syllabes de type « consonne-voyelle » issues des énoncés monosyllabiques et pluri-syllabiques (TABLE 1).

| | *Syllabes* | | | | *Cycles masticatoires* | | | |
|---|---|---|---|---|---|---|---|---|
| | *Session 1* | *Session 2* | *Session 3* | *Session 4* | *Session 1* | *Session 2* | *Session 3* | *Session 4* |
| *Participant 1* | *57* | *27* | *40* | *51* | *66* | *66* | *167* | *49* |
| *Participant 2* | *38* | *16* | *39* | *39* | *94* | *186* | *153* | *77* |
| *Participant 3* | *10* | *26* | *62* | *37* | *65* | *32* | *101* | *32* |
| *Participant 4* | *17* | *48* | *27* | *44* | *61* | *86* | *58* | *235* |
| ***Total*** | ***122*** | ***117*** | ***168*** | ***171*** | ***286*** | ***370*** | ***479*** | ***393*** |

TABLE 1 : Nombre de syllabes et de cycles masticatoires par participant pour chaque session

- Traitement statistiques des données :

En raison de la taille limitée de l'échantillon (n=4), les données de cette étude décrivent de manière qualitative l'évolution des patrons temporels mandibulaires observés au cours de la nutrition et de la parole. Le test non-paramétrique de Wilcoxon-Mann-Whitney a été utilisé pour comparer les durées syllabiques et les durées des cycles masticatoires sur l'ensemble de la période étudiée.

# 3 Résultats

## 3.1 Comparaison entre l'évolution de la durée syllabique moyenne et la durée moyenne d'un cycle masticatoire entre 8 et 14 mois

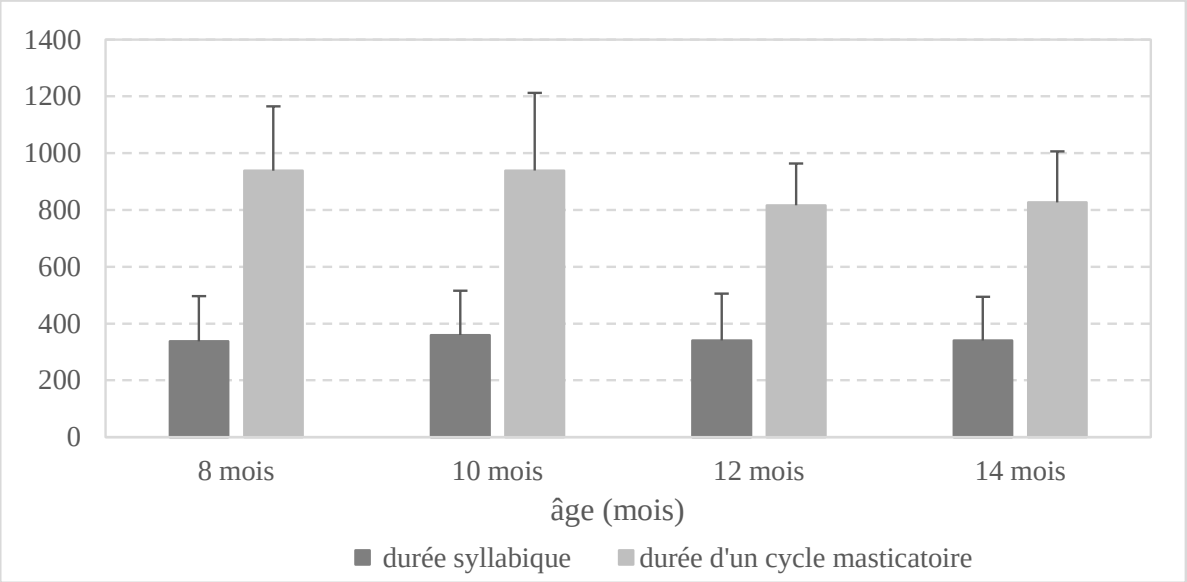



FIGURE 1 : Comparaison entre l'évolution de la durée syllabique moyenne et la durée moyenne d'un cycle masticatoire entre 8 et 14 mois

La durée syllabique moyenne reste relativement stable entre 8 mois (337 ms) et 14 mois (340 ms) (FIGURE 1) même si une légère augmentation est observée à 10 mois (360 ms). A contrario, une diminution de la durée moyenne du cycle masticatoire est observée entre 8 mois (921 ms) et 14 mois (855 ms) (FIGURE 1). Plus précisément, la durée du cycle masticatoire est stable entre l'âge de 8 mois et celui de 10 mois et une diminution semble s'amorcer à partir de 10 mois. Sur l'ensemble de la période d'observation, la durée d'un cycle masticatoire est significativement plus longue que la durée syllabique ($W=136$, $p< 0.0001$).

### 3.2 Comparaison entre l'évolution des coefficients de variation pour la durée syllabique moyenne et la durée moyenne d'un cycle masticatoire

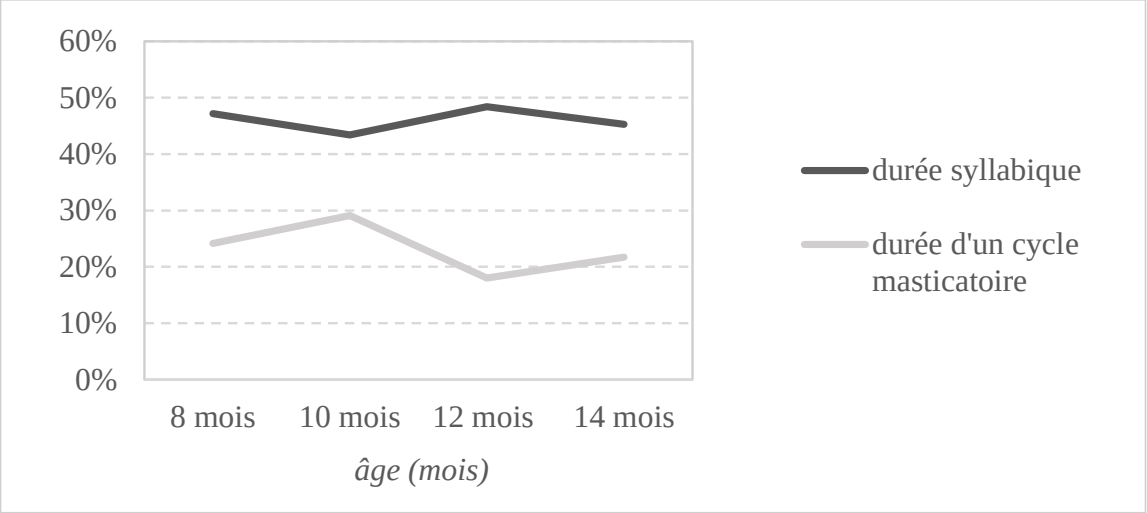



FIGURE 2 : Comparaison entre l'évolution des coefficients de variation pour la durée syllabique moyenne et la durée moyenne d'un cycle masticatoire entre 8 et 14 mois

Les coefficients de variation (écart-type/moyenne*100) mettent en évidence une variabilité plus importante des durées syllabiques (46,1 %) par rapport aux durées des cycles masticatoires (23,2 %) dès 8 mois. A 10 mois, un pic de variabilité est observé pour la durée moyenne d'un cycle masticatoire (29,1%) puis celle-ci diminue à 12 mois (18 %) et 14 mois (21,7 %). Un pic de variabilité est également observable pour la syllabe à 12 mois (48,4 %) (FIGURE 2).

# 4 Discussion

## 4.1 Un développement précoce non linéaire

En dépit de différences interindividuelles importantes et du nombre limité de participants, les résultats obtenus sont en accord avec les données de la littérature pour les durées syllabiques (Kent & Murray, 1982) et pour les durées des cycles masticatoires (Gisel, 1991) (FIGURE 1). La diminution de la durée syllabique et de celle du cycle masticatoire amorcée à partir de 10 mois mettent en évidence une évolution en deux étapes des durées des cycles oscillatoires mandibulaires pour les activités de parole et de nutrition qui pourrait illustrer l'amélioration des compétences articulatoires et masticatoires au cours du développement (Green et Nip, 2012). Cette non-linéarité a été mise en évidence à plusieurs reprises pour l'acquisition des compétences motrices (Green & Nip, 2012 ; Smith & Thelen, 2003). Selon ces auteurs, les processus à l'origine du développement moteur constitueraient un système complexe qui s'auto-organise et qui évolue par phases de stabilités et d'instabilités (e.g. accélérations, décélérations, plateaux) jusqu'à ce que le système devienne mature. Ces phases seraient alors issues de l'interaction entre des facteurs qui stimulent l'évolution des compétences motrices (e.g. développement cognitif, environnement) et des facteurs qui limitent cette évolution (e.g. contraintes bio-mécaniques) (Green & Nip, 2012 ; Smith & Thelen, 2003).

De ce fait, la trajectoire de ces patrons temporels observée entre 8 et 14 mois pourrait refléter, d'une part, la transition entre le babillage réduplicué et le babillage varié (Canault, 2007 ; Canault, 2011) et d'autre part, le développement des compétences masticatoires (Gisel, 1991).

## 4.2 La variabilité comme indice de développement des habiletés oro-motrices

Les résultats obtenus montrent des variabilités temporelles importantes pour la parole entre 8 et 14 mois (FIGURE 2). Le pic de variabilité observé à 10 mois pour la mastication pourrait être expliqué par la réorganisation motrice qui s'effectue au cours de cette période (FIGURE 2). En effet, si au début de la diversification alimentaire les mouvements de succion (« suckling ») et de mastication se chevauchent, c'est à partir de 10 mois que les mouvements masticatoires deviennent majoritairement présents (Gisel, 1991). Ce pic de variabilité semble émerger dans notre étude à l'âge de 12 mois pour la syllabe, alors que dans d'autres travaux il apparait autour de 10-11 mois (Canault, 2007 ; Canault, 2011). Cette période critique pourrait ainsi témoigner de l'amélioration des compétences oro-motrices (Green & Nip, 2012). La variation pourrait ainsi être utilisée par le bébé comme un moyen de se libérer des contraintes précoces pesant sur son système oro-moteur (MacNeilage 1998).

## 4.3 Spécificité précoce des activités de parole et de mastication ?

Dès l'âge de 8 mois, il existe une différence entre les durées des cycles oscillatoires mandibulaires observés pour la mastication et ceux observés pour la parole (FIGURE 1). Ces résultats préliminaires semblent ainsi nuancer le postulat émis par MacNeilage (1998) défendant que les cycles d'ingestion, de par leur activité rythmique stéréotypée, pourraient servir de précurseurs au développement de la parole même s'ils ne rejettent pas complétement l'hypothèse d'une interaction entre le développement des activités de parole et de mastication. En effet, la différenciation précoce des patrons temporels de ces deux activités pourrait reposer sur une activation musculaire spécifique à la tâche (Moore & Ruark, 1996), mais la rythmicité des cycles mandibulaires sous-jacents pourrait tout de même être générée par des structures cérébrales communes (Barlow & Estep, 2006). Ainsi, les mouvements rythmiques oscillatoires seraient contrôlés par un générateur central de pattern de mouvements (CPG) situés dans le tronc cérébral. Ces CPG sont à l'origine de tous les mouvements biphasiques rythmiques observés chez l'humain et sont composés de circuits simples ou complexes qui, modulés par des afférences sensorielles, agissent directement sur les motoneurones qui génèrent des patterns rythmiques de mouvements stéréotypés (Barlow & Estep, 2006). L'une des particularités des CPG réside dans le fait que certains des circuits qui les composent seraient capables de se réorganiser au cours du développement et produire des patterns temporels différents. Dans le cas présent, une sous-population de neurones située dans les CPG masticatoires pourrait être réorganisée pour générer des mouvements mandibulaires différents permettant l'émergence des oscillations rythmiques observées pour la parole (Grillner, 1982). Les patrons temporels rythmiques spécifiques obtenus dans cette étude pourraient ainsi illustrer le réaménagement précoce des CPG masticatoires pour permettre l'émergence de la parole. Il est à présent nécessaire d'examiner d'autres paramètres tels que l'évolution de l'amplitude ou de la fréquence oscillatoire mandibulaire pour comprendre quelles interactions peuvent exister entre le développement de la mastication et de la parole. Pour cela, une nouvelle étude transversale effectuée sur un plus grand nombre de sujets et complétée par des mesures cinématiques est actuellement en cours de réalisation.

# Remerciements

Ce projet a été financé par le Laboratoire d'Excellence ASLAN (« Advanced Studies on LANguage complexity », Université de Lyon), par le Fonds France-Canada pour la Recherche ainsi que par Mitacs Globalink- Campus France et le Programme Avenir Lyon Saint-Etienne.

# Références

BARLOW, S. M., & ESTEP, M. (2006). Central pattern generation and the motor infrastructure for suck, respiration, and speech. *Journal of Communication Disorders*, *39*(5), 366–380.

BICKLEY C., LINDBLOM B. & ROUGH L. (1986). Acoustic measures of rhythm in infants' babbling, or "All god's children got rhythm". Proceedings of the 12th International Congress on Acoustics, Toronto, A6-4.

CANAULT, M. (2007). *L'emergence du contrôle articulatoire au stade du babillage. Une étude acoustique et cinématique*. *Thèse de Doctorat,* Université Marc Bloch.

CANAULT, M., & LABOISSIERE, R. (2011). Le babillage et le développement des compétences articulatoires: indices temporels et moteurs. *Faits de langues,* (37).

DOLATA, J. K., DAVIS, B. L., & MACNEILAGE, P. F. (2008). Characteristics of the rhythmic organization of vocal babbling: Implications for an amodal linguistic rhythm. *Infant Behavior and Development*, *31*, 422–431.

GISEL, E. G. (1991). Effect of food texture on the development of chewing of children between six months and two years of age. *Developmental Medicine and Child Neurology, 33*, 69–79.

GOLDFIELD E.C. & WOLFF P.H. (2003) A dynamical systems perspective on infant action and its development. In Theories of infant development. Brenner J.G. & Slater A. (Eds.). Oxford, Blackwell Publishing, 3- 29.

GREEN, J. R., & NIP, I. S. B. (2012). Some organization principles in early speech development. *Speech Motor Control: New Developments in Basic and Applied Research*, 171–188.

GRILLNER, S. (1982). Possible analogies in the control of innate motor acts and theproduction of sound in speech. In: Speech Motor Control, edited by S. Grillner, P. Lindblom, J. Lubker, and A. Persson. 217-230.

JURGENS, U.(1998).Speech evolved from vocalization, not mastication. Commentaire à MacNeilage P.F. (1998). The Frame/Content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21, 519-520.

KENT R.D. & MURRAY A.D. (1982). Acoustic features of infant vocalic utterances at 3, 6, and 9 months. *Journal of the Acoustical Society of America*, 72 (2), 353-365.

LE REVEREND, B. J. D., EDELSON, L. R., & LORET, C. (2014). Anatomical, functional, physiological and behavioural aspects of the development of mastication in early childhood. *The British Journal of Nutrition*, *111*(3), 403–414.

MACNEILAGE, P. F. (1998). The frame/content theory of evolution of speech production. *The Behavioral and Brain Sciences*, *21*(4), 499-511-546.

MALAS, K., TRUDEAU, N., CHAGNON, M., & MCFARLAND, D. H. (2015). Feeding-swallowing difficulties in children later diagnosed with language impairment. *Developmental Medicine & Child Neurology*, *57*(9), 872–879.

MOORE, C. A, & RUARK, J. L. (1996). Does speech emerge from earlier appearing oral motor behaviors? *Journal of Speech and Hearing Research*, *39*(5), 1034–1047.

REILLY, S., SKUSE, D. H., & WOLKE, D. (2000). Schedule for Oral Motor Assessment (SOMA). Eastgardens. *New South Wales: Whurr*.
SMITH A. & GOFFMAN, L. (1998). Stability and patterning of speech movement sequences in children and adults. Journal of Speech, Language, and Hearing Research, 41, 18-30..
SMITH, L. B., & THELEN, E. (2003). Development as a dynamic system. *Trends in Cognitive Sciences*, *7*(8), 343–348.
SMITH, A., & ZELAZNIK, H. N. (2004). Development of functional synergies for speech motor coordination in childhood and adolescence. *Developmental Psychobiology*, *45*(1), 22–33.
STEEVE, R. W., & MOORE, C. A. (2009). Mandibular motor control during the early development of speech and nonspeech behaviors. *Journal of Speech, Language, and Hearing Research : 52*, 1530–1554.

# Effet de la position de la syllabe sur la réalisation acoustique des consonnes finales du thaï

Nicha YAMLAMAI, Thi Thuy Hien TRAN
GIPSA-lab, Département Parole et Cognition, UMR 5216
CNRS & Université Grenoble Alpes, BP 25, 38040 Grenoble Cedex 9, France
y.nicha@hotmail.com, thi-thuy-hien.tran@gipsa-lab.grenoble-inp.fr

## RESUME

De nombreuses études ont montré les différences dans les réalisations acoustiques des consonnes en fonction de leur position dans le mot. Ce travail compare les caractéristiques acoustiques des consonnes finales de syllabes identiques du thaï dans les différentes positions. Ces consonnes se trouvent soit devant une frontière de mot CVC#CVC, soit devant une frontière de syllabe à l'intérieur d'un mot dissyllabique CVC.CVC. L'objectif est de savoir si le type de frontière (inter-mots *vs*. intra-mot) a un effet sur la réalisation de ces consonnes. Les premiers résultats montrent que non seulement la durée de la consonne finale mais aussi celle de la voyelle précédente et celle de la syllabe sont significativement influencées par le type de frontière. D'autres paramètres acoustiques sont en cours d'analyse. Il s'agit d'une étape préliminaire à une étude ayant pour but de comprendre les caractéristiques de la langue source (le thaï) susceptibles d'être des éléments pouvant empêcher l'acquisition des consonnes finales de la langue cible (le français).

## ABSTRACT

**Effects of syllable position on the acoustic realization of Thai final consonants**

Various studies have shown differences in the acoustic realization of consonants as a function of their position in the word. This paper presents an acoustical study of Thai final consonants in identical syllables, in relation to the types of boundary (before a word boundary CVC#CVC or before a syllable boundary CVC.CVC). The aim of this study is to understand if the boundary type (inter-words *vs*. intra-word) has an effect on the realization of these final consonants. The very first results show that not only the duration of the final consonants but also those of the preceding vowel and of the syllable are significantly affected by the types of boundary. Further acoustical parameters are currently being analyzed. It is a preliminary step to study and understand which characteristics of the source language potentially acting as a phonological sieve to the acquisition of French final consonants.

MOTS-CLES : étude acoustique, consonnes finales, frontière syllabique, frontière de mots, thaï
KEYWORDS : acoustic study, final consonants, syllabic boundary, word boundary, Thai

# 1 Introduction

De nombreuses études ont mis en évidence des différences dans les productions de consonnes en fonction de leur position dans la syllabe (attaque *vs*. coda) (Lindblom, 1983 ; Browman et Goldstein, 1995) ou dans le domaine prosodique, par ex. énoncé (Fougeron et Keating, 1997). Dans les langues du monde, les structures permettant uniquement une (des) consonne(s) en attaque


10.21437/JEP.2018-18

sont plus nombreuses que celles avec attaque et coda pleines, et dépassent de loin celles n'autorisant que la (les) consonne(s) en coda (Redford et Diehl, 1999 ; Rousset, 2004). Il a été également montré que les consonnes finales sont défavorisées car produites avec moins de contact lingual et moins de pression linguopalatale (Browman et Goldstein, 1995; Fougeron, 1999), elles sont moins faciles à identifier et/ou plus difficiles à produire que les initiales (Redford et Diehl, 1999, pp. 1555).

Des effets de la position intra-mot et intra-syllabe ont été trouvés sur le timing et la coarticulation des séquences de plosives (Recasens *et al.*, 1993 ; Zsiga, 1994 ; Byrd, 1996 ; Davidson, 2007 ; Hoole et *al.*, 2012). Il a été montré que les groupes de consonnes tautosyllabiques, que ce soit en attaque (#CC) ou en coda (CC#) sont produits avec un degré de coarticulation plus important que les séquences de consonnes hétérosyllabiques (C#C) (Davidson, 2007).

Que se passe-t-il alors pour une langue comme le thaï où les groupes de consonnes tautosyllabiques sont interdits en coda (*CVCC) et où les consonnes finales se trouvent seulement en séquences hétérosyllabiques soit à la frontière de mot CVC#CVC, soit à la frontière de syllabe à l'intérieur d'un mot CVC.CVC ? Étant donnée cette contrainte phonotactique, existe-t-il un effet du type de frontière (inter-mots *vs*. intra-mot) sur la production des consonnes finales, qui proviendrait d'un degré de coarticulation plus fort (plus de chevauchement) entre les consonnes à la frontière de syllabe à l'intérieur d'un mot (C.C) qu'à la frontière de mots (C#C) ?

Le thaï, appartenant à la famille tai-kadaï, est une langue tonale et isolante où les mots sont invariables quelle que soit leur fonction grammaticale (Iwasaki et Ingkaphirom, 2005). Les mots monosyllabiques et dissyllabiques occupent une place majoritaire dans le lexique du thaï (respectivement 41,37 % et 40,35 % des unités lexicales) (Yamlamai, 2017). La structure syllabique peut se présenter selon le modèle $C_1(C_2)V(C_3)$ (entre parenthèses, les constituants facultatifs) (Abramson, 1962). Le type syllabique le plus recruté est CVC, cette structure constitue à elle seule presque 65 % des syllabes de la langue (Rousset, 2004). Une autre particularité du thaï est l'inventaire très restreint des consonnes en coda. A part les glides /w j/ et la glottale /ʔ/, seules six occlusives /p t k m n ŋ/ peuvent être en coda (Iwasaki et Ingkaphirom, 2005). De plus, les plosives sourdes /p t k/ sont non relâchées en position finale ne générant pas un bruit d'explosion audible après la partie d'occlusion (Tingsabadh et Abramson, 1999 ; Tsukada, 2004).

En tenant compte des caractéristiques du thaï mentionnées ci-dessus, l'objectif de l'expérience présentée ici est d'étudier sur le plan acoustique les consonnes plosives /p t k/ et nasales /m n ŋ/ du thaï en comparant leur réalisation en fonction de leur position : (1) en position de coda $C_{2\#}$ dans les mots simples $C_1V\mathbf{C_{2\#}}$ ; (2) en position de coda $\mathbf{C_{2\sigma}}$ de la première syllabe des mots dissyllabiques de structure $C_1V\mathbf{C_{2\sigma}}.C_3VC_4$.

Ce travail s'inscrit dans un cadre plus général d'acquisition des langues secondes et sert de base à des travaux ultérieurs sur l'apprentissage des consonnes finales du français par les apprenants thaïlandais. Il s'agit d'une étape préliminaire à la compréhension des caractéristiques de la langue source (thaï) susceptibles d'être des éléments du crible phonologique pouvant gêner l'acquisition des consonnes finales de la langue cible (français).

# 2 Méthodologie

## 2.1 Constitution du corpus

Les mots monosyllabiques $C_1VC_2$ et dissyllabiques $C_1VC_2.C_3VC_4$, dont la syllabe $C_1VC_2$ est identique, ont été sélectionnés pour la constitution du corpus. Rappelons que le thaï est une langue polytonale qui possède cinq tons phonologiques. Le dialecte du centre comporte trois tons ponctuels (tons moyen, bas et haut) et deux tons modulés (tons descendant et montant) (Abramson, 1972 ; Iwasaki et Ingkaphirom, 2005). Afin de faciliter la segmentation des

paramètres acoustiques, notamment la fréquence fondamentale, un ton ponctuel (préférable à un ton modulé) a été choisi pour l'étude. Parmi les trois tons ponctuels, le ton haut a été sélectionné pour les syllabes cibles ($C_1VC_2$) en raison de l'inexistence du ton moyen dans une syllabe terminée par des plosives. En cas de mots composés dissyllabiques, le même ton haut a été gardé pour la deuxième syllabe à plosive finale parmi /p t k/, le ton moyen a été choisi pour les syllabes à sonante finale parmi /m n ŋ w j/. Bien que le ton bas soit également ponctuel, les deux tons moyen et haut ont été sélectionnés pour la deuxième syllabe de mots composés, en raison de la proximité de ces tons dans l'espace tonal en terme de hauteur fréquentielle (Abramson, 1962). Nous avons sélectionné la voyelle brève /a/ qui fournit un meilleur contraste entre segments consonantiques. À noter que les voyelles du thaï peuvent s'opposer sur le trait de longueur. Cependant, les voyelles brèves sont plus fréquentes sur le plan lexical, les mots contenant la voyelle longue /aː/ et le ton haut sont majoritairement des emprunts à l'anglais. Pour les mots dissyllabiques, les consonnes $C_3$ ont été choisies de manière à ce que toute consonne sourde /p t k/ en $C_2$ soit suivie d'une consonne sonore en $C_3$, et inversement, toute consonne sonore /m n ŋ/ en $C_2$ soit suivie en $C_3$ d'une consonne sourde. La table 1 présente 60 stimuli (30 mots monosyllabiques, 30 mots composés dissyllabiques) qui répondent à l'ensemble de ces critères.

| Mot | Signification | Mot | Signification |
|---|---|---|---|
| kʰráŋ | fois | nán | cela |
| kʰráŋ.kʰraːw | de temps en temps | nán.lɛː | voilà |
| kʰát | sélectionner | náp | compter |
| kʰát.ŋáːŋ | être aux prises avec quelqu'un | náp.wan | de jour en jour |
| kʰám | soutenir | pʰát | éventail |
| kʰám.kʰɔː | obliger | pʰát.jót | éventail qui désigne le grade du moine |
| ŋát | ouvrir en faisant levier | fák | courge |
| ŋát.ŋɛ́ʔ | forcer une ouverture | fák.mɛ́w | chayote |
| tɕʰák | retirer | mák | souvent |
| tɕʰák.nam | inciter | mák.nój | sans ambition |
| tɕʰán | étage | ják | hausser (sourcils ou épaules) |
| tɕʰán.tɕʰɤːŋ | stratégie | ják.jáːj | déplacer |
| tɕʰám | être contusionné | Jáŋ | empêcher |
| tɕʰám.tɕaj | être blessé | jáŋ.kʰít | réfléchir avant d'agir |
| sák | laver des vêtements | jáp | froissé |
| sák.láːŋ | laver | jáp.jáŋ | entraver |
| sát | se projeter | rák | aimer |
| sát.nám | jeter de l'eau lors du mariage | rák.jom | amulette |
| sáp | absorber | ráŋ | retenir |
| sáp.naj | sous-vêtement | ráŋ.tʰáːj | dernier rang |
| sám | répéter | lák | voler |
| sám.tɤːm | en rajouter | lák.jím | fossette |
| tʰáŋ | entier | lát | prendre un raccourci |
| tʰáŋ.puaŋ | tout | lát.lɔ́ʔ | contourner |
| tʰáp | poser sur quelque chose | láp | secret |
| tʰáp.lék | une sorte d'insecte | láp.líː | caché |
| nák | très | wát | temple |
| nák.bin | pilote | wát.waː | temple (réduplication) |
| nát | rendez-vous | wáp | soudainement |
| nát.nɛ́ʔ | fixer un rendez-vous | wáp.waːw | scintiller |

TABLE 1 : Liste des 60 stimuli pour l'étude acoustique des consonnes finales du thaï.

Ces mots sont insérés dans une phrase porteuse : « พูดคำว่า [mot cible] ไวๆ » /pʰûːt kʰam wâː [mot cible] waj waj/ ou « dire [mot cible] vite ». Le corpus est constitué de 4 répétitions des 60 mots insérés dans les phrases porteuses et présentées dans un ordre aléatoire. Ces 240 phrases sont lues par une locutrice et un locuteur natifs thaïlandais originaires de Bangkok et parlant tous deux le thaï standard (variété du Centre). Le corpus a été enregistré dans la chambre sourde du Département Parole et Cognition de GIPSA-lab avec enregistreur numérique Marantz PMD 670, Micro AKG C1000S à directivité cardioïde. L'enregistrement a été numérisé à 44,1 kHz sous format WAV, avec une résolution de 16 bits.

## 2.2 Segmentation et analyse des données

La segmentation des 480 stimuli a été réalisée manuellement avec Praat. À l'aide de scripts, nous avons extrait de manière semi-automatique la durée des segments. Nous nous intéressons à la durée des consonnes finales /p t k m n ŋ/, à la durée de la voyelle /a/ qui précède la consonne cible, à la durée de la syllabe cible (désormais $S_1$) : $\underline{C_1}/a/\underline{C_{2\#}}$ et $\underline{C_1}/a/\underline{C_{2\sigma}}.C_3VC_4$ ainsi qu'au VOT des plosives finales en cas de relâchement du burst. En raison de la fréquence des plosives non relâchées en coda, les paramètres acoustiques concernant l'évolution temporelle de la transition entre /a/ et plosive finale à partir de 50 % jusqu'à 90 % de la durée de la voyelle (fréquence fondamentale, trois premiers formants, intensité) ont été extraits et seront prochainement analysés. Dans le cadre de ce travail, seules les mesures concernant le paramètre de durée ont été analysées.

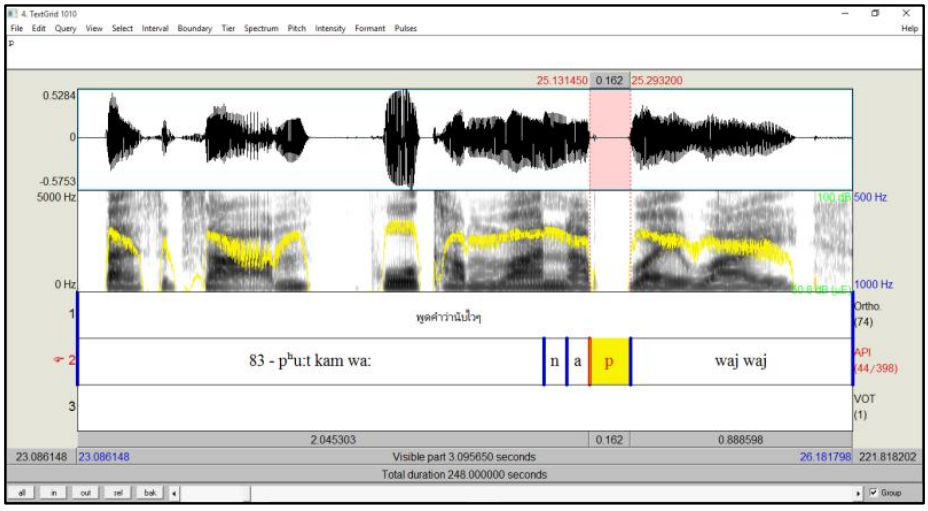



FIGURE 1 : Exemple de segmentation sur le logiciel Praat.

Toutes les mesures de durées ont été normalisées en fonction du débit de parole en divisant la durée absolue des segments par la valeur du débit local. Le débit local (interprété comme le nombre de syllabes par seconde) est égal au nombre de syllabes divisé par la durée absolue du mot en seconde. Nous avons utilisé le logiciel SPSS© (Statistical Package for the Social Sciences) pour les analyses statistiques de ces données normalisées. Des ANOVA ont été réalisés pour rechercher l'effet de la position $C_{2\#}$ ou $C_{2\sigma}$ sur la durée de la consonne finale ($C_2$), la durée de la voyelle et la durée de la syllabe $S_1$.

# 3 Résultats

## 3.1 Durée de la consonne finale

Les résultats montrent qu'en général, les consonnes plus longues se trouvent en finale de mot $C_{2\#}$ alors que les réalisations plus brèves, quelle que soit la consonne, sont en coda de syllabe à l'intérieur d'un mot composé ($C_{2\sigma}$) (cf. figure 2).

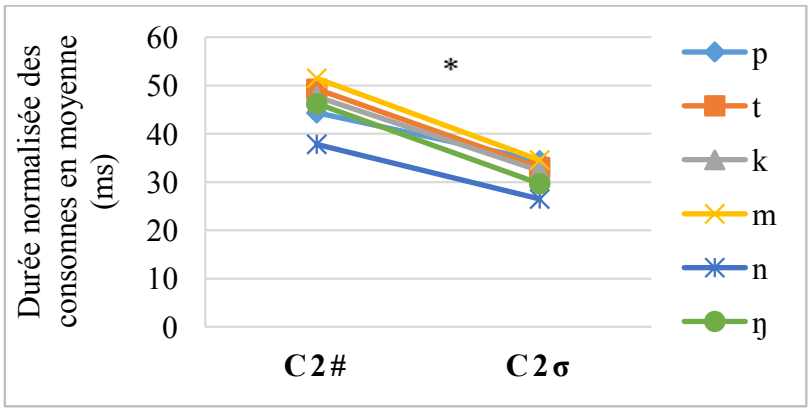



FIGURE 2 : Valeur moyenne des durées normalisées des consonnes en fonction du type de frontière.

Un effet significatif est observé globalement en inter-sujet pour toutes les consonnes [$F(1, 468) = 71{,}079$ ; $p < 0{,}05$]. En fonction du mode d'articulation, le type de frontière (inter-mots *vs*. intra-mot) a aussi un effet significatif sur la différence de durée des plosives [$F(1, 334) = 56{,}943$ ; $p < 0{,}05$] et des nasales [$F(1, 142) = 35{,}540$ ; $p < 0{,}05$]. Les plosives et les nasales sont significativement plus longues en coda de mot ($C_{2\#}$) qu'en coda de syllabe ($C_{2\sigma}$). Cette différence est significative est observée pour les bilabiales [$F(1, 142) = 16{,}900$ ; $p < 0{,}05$], les coronales [$F(1, 142) = 30{,}008$ ; $p < 0{,}05$] et les vélaires [$F(1, 190) = 47{,}213$ ; $p < 0{,}05$].

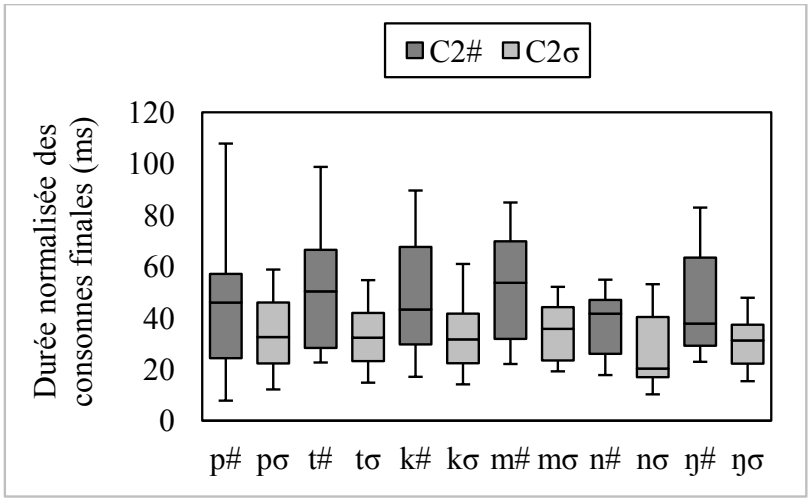



FIGURE 3 : Durée normalisée des consonnes en fonction de leur position ($C_{2\#}$ *vs*. $C_{2\sigma}$).

Aucune différence significative n'est observée en intra-sujet entre plosives et nasales [$F(1, 478) = 1{,}037$ ; $p = 0{,}309$], ni entre les lieux d'articulation (bilabial *vs*. coronal *vs*. vélaire) [$F(2, 477) = 0{,}282$ ; $p = 0{,}754$]. Un effet du locuteur est observé pour la durée de la consonne [$F(1, 476) = 918{,}218$ ; $p < 0{,}05$]. Quelle que soit sa position, la durée normalisée en moyenne de la consonne

finale est plus longue chez la locutrice que chez le locuteur (64 ms *vs*. 30 ms en $C_{2\#}$ et 41,4 ms *vs*. 23,4 ms en $C_{2\sigma}$).

Un débit de parole plus lent est relevé pour le locuteur féminin (~ 3 syllabes par seconde) et pour le locuteur masculin (~ 4 syllabes par seconde). On trouve donc un effet de l'interaction entre le facteur « locuteur » et la position de la consonne finale [$F(1, 476) = 88{,}271$ ; $p < 0{,}05$]. En intra-sujet, la différence selon la position reste significative pour toutes les consonnes seulement chez le locuteur féminin ($p < 0{,}05$ pour /p/, /t/, /k/, /m/ et /ŋ/ et $p = 0{,}029$ pour /n/). Par contre, il n'y a pas de différence significative observée de la durée en fonction de la position pour /p/ [$F(1, 46) = 0{,}168$ ; $p = 0{,}684$], ni pour /n/ [$F(1, 14) = 3{,}769$ ; $p = 0{,}78$] chez le sujet masculin, sans doute en raison de la grande variabilité de durées mesurées chez ce locuteur (cf. figure 3).

## 3.2 Durée de la voyelle

La voyelle /a/ est significativement plus longue dans la structure $C/a/C_{2\#}$ que dans la structure $C/a/C_{2\sigma}$ [$F(1, 468) = 192{,}442$ ; $p < 0{,}05$] (cf. figure 4). La différence reste globalement significative que ce soit selon le mode [$F(1, 468) = 6{,}122$ ; $p = 0{,}014$] ou selon le lieu d'articulation [$F(2, 477) = 11{,}701$ ; $p < 0{,}05$]. De même, la durée de /a/ est significativement plus longue dans la structure $C/a/C_{2\#}$, qu'elle soit suivie d'une plosive [$F(1, 334) = 180{,}234$ ; $p < 0{,}05$] ou d'une nasale [$F(1, 142) = 49{,}032$ ; $p < 0{,}05$]. Par ailleurs, la durée de /a/ change de manière significative en fonction du lieu d'articulation des consonnes finales : elle est significativement plus longue quand elle est suivie d'une vélaire (27 ms en moyenne) que d'une bilabiale (24 ms en moyenne) ou coronale (24 ms en moyenne) ($p < 0{,}05$). La durée la plus importante est relevée lorsque /a/ est suivie de la nasale vélaire, quelle que soit la position de la consonne finale : en moyenne 36 ms (/ŋ/ en $C_{2\#}$) et 26 ms (/ŋ/ en $C_{2\sigma}$). Aucune différence n'est observée lorsque /a/ est suivie d'une bilabiale ou d'une coronale ($p = 0{,}608$).

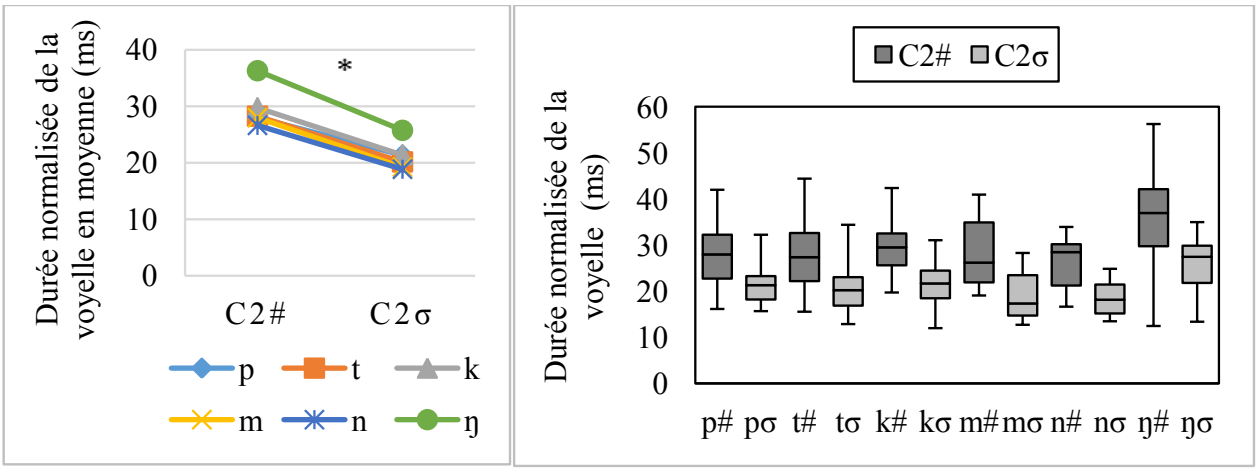



FIGURE 4 : Durée normalisée de la voyelle /a/ devant plosive et nasale selon le type de frontière.

## 3.3 Durée de la première syllabe

Rappelons qu'il s'agit toujours de la même syllabe C/a/C qui est soit un monosyllabe soit la première syllabe d'un composé. Nous observons la même tendance que pour les deux paramètres précédemment étudiés : la durée de la $S_1$ en mot monosyllabique est significativement plus longue qu'en mot dissyllabique [$F(1, 468) = 85{,}187$ ; $p < 0{,}05$] (cf. figure 5). La durée de la $S_1$ est différente significativement, selon le mode d'articulation des consonnes finales, que ce soit des plosives [$F(1, 334) = 72{,}345$ ; $p < 0{,}05$] ou des nasales [$F(1, 142) = 33{,}539$ ; $p < 0{,}05$]. Le même effet significatif est observé en fonction du lieu d'articulation des bilabiales [$F(1, 142) = 22{,}948$ ;

p < 0,05], des coronales [F(1, 142) = 34,994 ; p < 0,05] et des vélaires [F(1, 190) = 48,990 ; p < 0,05].

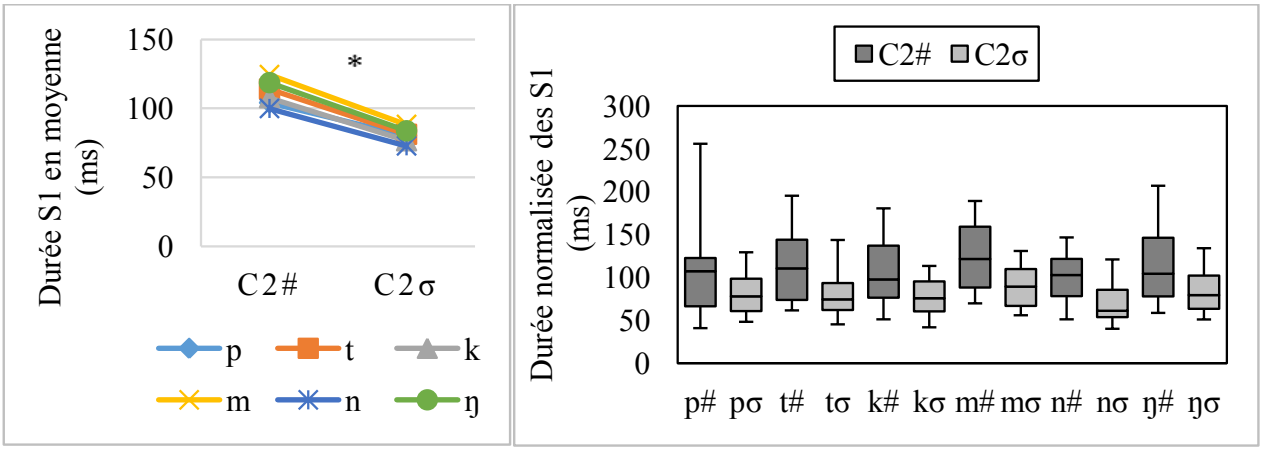



FIGURE 5 : Durée normalisée de la première syllabe en fonction du type de position.

# 4 Discussion et conclusion

Les résultats confirment l'existence de durées acoustiques différentes pour les consonnes finales du thaï en fonction du type de frontière (inter-mots *vs*. intra-mot). Ces différences sont relevées non seulement au niveau de la consonne finale, mais aussi de la voyelle /a/ et de la syllabe. Nos analyses montrent qu'une consonne est plus longue en finale de mot ($C_{2\#}$) qu'en finale de syllabe à l'intérieur d'un mot composé ($C_{2\sigma}$). De la même manière, la durée de la voyelle /a/ qui précède $C_{2\#}$ est plus longue que lorsque suivie par $C_{2\sigma}$. Ce résultat est valable pour les plosives comme pour les nasales. De plus, la durée d'une syllabe dépend du fait qu'elle constitue un mot monosyllabique ou une syllabe à l'intérieur d'un mot composé : elle est plus longue dans le premier cas. Ces résultats supposent clairement des degrés différents de coarticulation entre les segments en fonction du type de frontière syllabique.

Ce résultat est comparé à celui obtenu pour le vietnamien par Tran (2011). Dans cette langue, à part /w j/, seules les plosives /p t k/ et les nasales /m n ŋ/ appartiennent à l'inventaire en coda, comme en thaï. Les plosives du vietnamien possèdent également la caractéristique particulière du non relâchement. En vietnamien, le même effet de frontière sur la durée des segments a été observé. Les segments de la rime sont plus longs quand ils précèdent une frontière inter-mots qu'intra-mot. Ces résultats sont cohérents avec les travaux antérieurs qui montrent que la prononciation d'un phonème peut être influencée par la position dans le mot : début, interne, finale (Keating, Wright et Zhang, 1999).

Nous avons montré également que la voyelle /a/ est plus longue devant les vélaires que devant d'autres lieux d'articulation. Ce résultat est intéressant pour l'étude des consonnes finales non relâchées en thaï. Si les noyaux vocaliques contiennent dans leur partie finale des éléments acoustiques de transition vers la réalisation de la cible consonantique qui suit (Tran, 2011), notre étude montre également que la durée du segment vocalique comporte des indices sur le lieu d'articulation des consonnes qui les suivent.

Concernant les plosives finales du thaï, elles sont généralement décrites comme non relâchées en raison de l'absence de burst après la partie d'occlusion. Ce travail confirme la caractéristique non-relâchée des plosives finales du thaï déjà remarquée dans la littérature (Tingsabadh et Abramson, 1999 ; Tsukada, 2004). En effet, dans la plupart des cas, aucune trace du bruit d'explosion des plosives n'est attestée sur le spectrogramme, quel que soit le type de frontière syllabique (cf. figures 6 et 7).

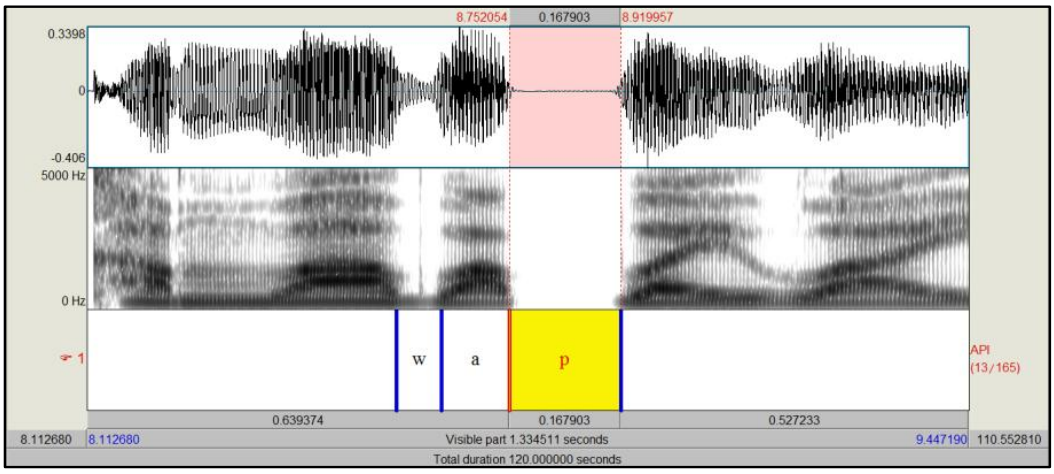



FIGURE 6 : Sonagramme de la plosive finale [p] du monosyllabe « วับ » [wáp] ou « soudainement » (locuteur féminin).

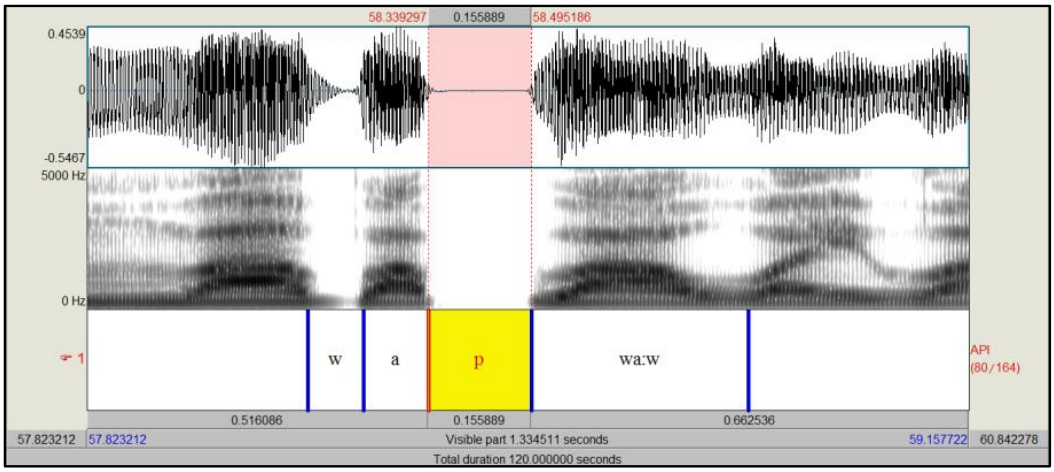



FIGURE 7 : Sonagramme de la plosive finale [p] en position intersyllabique du mot composé « วับวาว » [wáp.wa:w] ou « scintiller » (locuteur féminin).

Un cas de burst a été relevé pour [k] dans le mot [sák.lá:ŋ] et ce pour les deux sujets. Le burst de [k] est de durée plus brève et d'énergie plus faible (~ 60 dB en moyenne) par rapport à un [k] initial (~ 70 dB). Ce fait résulte probablement du mode articulatoire et de l'écoulement continu du [l] qui suit le [k], ainsi que de l'existence d'un groupe consonantique [kl] en initial. Ce cas ne représente que 2,38 % des productions analysées dans cette étude.

De manière générale, les résultats de cette étude révèlent que le type de frontière (inter-mots *vs*. intra-mot) influence la réalisation acoustique de la durée des consonnes finales, de la voyelle précédente et même de la syllabe, et suggèrent des degrés différents de coarticulation entre les segments. Les segments sont plus longs s'ils sont suivis d'une frontière inter-mots que s'ils sont suivis d'une frontière intra-mot. L'étude confirme également la tendance au non relâchement pour les plosives finales du thaï.

Les résultats que nous obtenons ici sur les différentes caractéristiques des consonnes finales du thaï en fonction du type de frontière syllabique peuvent constituer les éléments du crible phonologique de la langue source capables de gêner l'acquisition des consonnes finales du français par des apprenants thaïlandais. Nous complétons actuellement ce travail avec une étude perceptive et une analyse d'autres paramètres acoustiques (transitions formantiques, fréquence fondamentale, intensité).

# Remerciements

À Ratthapat Charoenwutipong et Panupan Junfeung, locuteurs natifs du thaï, pour leur participation à l'étude.


# Références


ABRAMSON, A. S. (1962). *The vowels and tones of standard Thai: Acoustical measurements and experiments*. Bloomington: Indiana U. Research Center in Anthropology, Folklore, and Linguistics, Pub. 20.

ABRAMSON, A. S. (1972). Word-final stops in Thai. In *Tai Phonetics and Phonology*. Bangkok : Central Institute of English Language, pp. 1-7.

BROWMAN, C. P., GOLDSTEIN, L. (1995). Gestural syllable position effects in American English. In *Producing Speech: Contemporary Issues: for Katherine Safford Harris* (Ed R. Bell-Berti). Woodbury, New York : AIP Press, pp. 19-33.

BYRD, D. (1996). Influences on articulatory timing in consonant sequences. *Journal of Phonetic* 24, 209-244.

DAVIDSON, L. (2007). Coarticulation in contrastive Russian stop sequences. In *Proceedings of the 16th International Congress of Phonetic Sciences*, Saabrüken, 417-420.

FOUGERON, C, KEATING, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America* 101(6), 3728-3740.

FOUGERON, C. (1999). Articulatory properties of initial segments in several prosodic constituents in French. *UCLA Working Papers in Phonetics* 97, 74-99.

HOOLE, P., BOMBIEN, L., MOOSHAMMER, C., POUPLIER, M., KÜHNERT, B. (eds.) (2012). *Consonant Clusters and Structural Complexity*. Berlin & New York : Mouton de Gruyter.

IWASAKI, S., INGKAPHIROM, P. (2005). *A Reference Grammar of Thai*. Cambridge : Cambridge University Press.

KEATING, P., WRIGHT, R., ZHANG, J. (1999). Word-level asymmetries in consonant articulation. *UCLA Working Papers in Phonetics* 97, 157‑ 173.

LINDBLOM, B. (1983). Economy of speech and gestures. In *The Production of Speech* (Ed P. F. MacNeilage). Berlin : Springer-Verlag, pp. 217-246.

RECASENS, D., FONTDEVILA, J., PALLARÈS, M. D., SOLANAS, A. (1993). An electropalatographic study of stop consonant clusters. *Speech Communication* 12, 335-355.

REDFORD, M., DIEHL, R. (1999). The relative perceptual distinctiveness of initial and final consonants in CVC syllables. *The Journal of the Acoustical Society of America* 106(3), 1555-1565.

ROUSSET, I. (2004). *Structures syllabiques et lexicales des langues du monde. Données, typologiques, tendances universelles et contraintes substantielles* (Thèse de Doctorat en Science du Langage). Université Stendhal, Grenoble 3.

TINGSABADH, K., ABRAMSON, A. S. (1999). Thai Final Stops: Cross-Language Perception. *Phonetica* 56, 111-122.

TRAN, T. T. H. (2011). *Processus d'acquisition des clusters et autres séquences de consonnes en langue seconde : de l'analyse acoustico-perceptive des séquences consonantiques du vietnamien à l'analyse de la perception et production des clusters du français par des apprenants vietnamiens du FLE* (Thèse de Doctorat en Science du Langage et Français Langue Etrangère). Université de Grenoble.

TSUKADA, K. (2004). Cross-language perception of final stops in Thai and English : a comparison of native and non-native listeners. In *Proceedings of the Tenth Australian International Conference on Speech Science & Technology, in Ryde, NSW*, 563-568.

YAMLAMAI, N. (2017). *Étude phonémique et acoustique des consonnes finales du thaï* (Mémoire de Master 1 en Sciences du Langage). Université Grenoble Alpes.

ZSIGA, E. C. (1994). Acoustic evidence for gestural overlap in consonant sequences. *Journal of Phonetic* 22, 121-140.

# Effets de l'orthographe dans la prononciation du français L2

Fabián Santiago[1,2,3]
(1) UMR 7023 CNRS, SFL, Université Paris VIII, 75017, Paris, France
(2) UMR 7018 CNRS, LPP, Université Sorbonne-Nouvelle, 75005 Paris, France
(3) UMR 7110 CNRS, Université Paris Diderot, 75013, Paris, France
fabian.santiago-vargas@univ-paris8.fr

## RÉSUMÉ

Nous étudions les effets de deux types de tâches (imitation *vs* lecture oralisée), le rôle de l'orthographe et le niveau de langue dans la prononciation du français L2 chez 27 étudiants hispanophones. Pour évaluer la prononciation en L2, nous analysons les distances de Levenshtein entre les transcriptions phonétiques de la forme canonique et celles de la forme produite par les participants. Cette analyse concerne 3,6k mots produits par les participants. Les résultats montrent que les erreurs de prononciation augmentent de 10% dans la lecture oralisée par rapport à la tâche d'imitation. Nous trouvons que la compétence phonique ne change pas en fonction du niveau de langue. Nous faisons l'hypothèse que ces erreurs sont dues, outre le transfert phonétique-phonologie de la L1 sur la L2, aux interférences négatives de l'orthographe (associations erronées entre graphie et son). Nous discutons de l'impact négatif que peut avoir l'input écrit dans l'acquisition de la phonétique/phonologie en L2, facteur souvent ignoré dans les recherches en phonologie.

## ABSTRACT

We report the effects of two types of tasks (imitation *vs* reading), orthography and proficiency level on the pronunciation accuracy in the speech of 27 Mexican Spanish learners of L2 French. We calculate Levenshtein distances between the phonetic transcriptions of the canonical pronunciation in L1 French and the actual learners' productions as an evaluation of the pronunciation accuracy. The analysis was carried out on 3.6k words produced by the participants. Results show that pronunciation error rates increase by 10% in the reading task with respect to the imitation task. We did not find any effect of the proficiency level. We propose that these errors are due, apart from the L1 phonological/phonetic transfer, also to the negative L1 transfer of letter-to-sound correspondences on the L2. We discuss the negative effects of written input on the acquisition of L2 phonology (a factor that is neglected in current L2 phonology models).

MOTS-CLÉS : prononciation en FLE, erreurs de prononciation en FLE, orthographe du français.
KEYWORDS: pronunciation in L2 French, pronunciation errors in L2 French, French orthography.

# 1 Introduction

Les apprenants adultes d'une L2 en contexte formel (apprentissage de la langue cible dans un pays où la L2 n'est pas une langue officielle/parlée) sont souvent confrontés à l'input écrit depuis le début de l'acquisition. Il est donc fréquent que les apprenants/enseignants emploient, outre l'input auditif ou les exercices d'entraînement articulatoire, le support de l'écrit pour apprendre et enseigner la


10.21437/JEP.2018-19

prononciation des nouveaux sons. En l'occurrence, les règles d'association entre les graphies et les sons dans la L2 sont souvent utilisées pour apprendre le nouveau système sonore de la L2.

L'apprenant adulte d'une L2 peut être influencé, non seulement par la phonétique et la phonologie de sa L1, mais également par les effets de l'orthographe de la L1 et de la L2 elle-même. Les apprenants sont donc confrontés à deux types de problèmes : la difficulté de prononcer et percevoir les nouveaux sons de la langue cible et la difficulté d'éviter une prononciation orthographique ou des associations erronées entre les graphies et les sons (cf. Detey *et al*. 2005, Young-Sholten 2002).

L'apprenant doit surmonter les inconsistances des rapports graphèmes-phonèmes en français L2, langue avec un système d'orthographe opaque (cf. Catach 2004) : (i) un phonème peut correspondre à différentes configuration graphiques (/o/ peut correspondre aux suites <o>, <au> ou <eau>), (ii) une graphie peut correspondre à différents sons (<s> se prononce [s] dans le mot *sel* mais [z] dans le mot *base*), (iii) une graphie n'est pas prononcée (<t> dans le mot *chat*), etc. L'objectif de cette étude est d'examiner les effets de la lecture oralisée et de l'orthographe dans la prononciation du français L2 chez les apprenants hispanophones du français L2 au Mexique. Notre intérêt sera d'examiner l'impact que peuvent avoir les systèmes d'orthographe opaque, comme c'est le cas du français, dans l'acquisition de l'émergence de l'interlangue phonologique d'une L2.

# 2 Le rôle de l'orthographe dans la prononciation en L2

La place qu'occupe l'orthographe dans l'apprentissage de la prononciation a été largement négligée dans les nombreux modèles d'acquisition de la phonologie/phonétique des L2 (cf. le *Speech Learning Model* (Flege 1995), le *Perceptual Assimilation Model* (Best 1995) ou encore le *Second Language Linguistic Perception Model* (Escudero 2005), entre autres). Il faut noter que ce n'est pas le cas de l'influence de l'orthographe dans la prononciation en L1 chez les monolingues, car différentes études en psycholinguistique ont examiné l'influence de l'orthographe dans la reconnaissance des phonèmes/mots (Frauenfelder *et al*. 1990) et dans la variation phonétique (Chevrot & Malderez 1999)

L'influence de l'orthographe dans la prononciation en L2 n'a attiré l'attention des chercheurs que depuis quelques années. Les études consacrées à ce sujet se classent en deux domaines. Le premier concerne l'interaction entre les représentations orthographiques de la L1/L2 dans le développement de l'interphonologie au niveau de la production. C'est le cas de l'influence des graphies dans la prononciation des approximantes en espagnol L2 par des anglophones : lorsque la suite en espagnol *la vaca* (la vache) est prononcée [la.'va.ka] au lieu de [la.'βa.ka] dû à la présence de la graphie <v> (Zampini 1994). Un autre exemple est celui de l'influence des graphies silencieuses dans l'insertion des voyelles épenthétiques en anglais L2 par des lusophones du Brésil : lorsque le mot anglais *tape* est prononcé ['tej.pi] ou ['tej.pə] dû à la présence de la lettre finale <e> (Silveira 2007).

Le deuxième domaine s'intéresse à évaluer les effets de l'orthographe dans la perception/imitation en L2. Les travaux de Bassetti (2017) et Detey *et al*. (2005) vont dans ce sens. Ces séries de travaux ont mis en évidence que les effets de l'orthographe sont, en quelque sorte, durables dans la production orale, si bien que seulement les effets des graphies peuvent expliquer les erreurs en imitation ou reconnaissance des mots dans la langue cible. Ainsi, certaines erreurs de prononciation seraient motivées par des interférences de l'orthographe de la L1-L2, même lorsque les apprenants articulent la L2 en dehors de la présence de l'écrit. Tout ceci nous amène à considérer que, outre l'impact de la L1, l'orthographe peut avoir des effets sur l'acquisition de la phonologie/phonétique en L2. Bien que pour la plupart des théories en phonologie de L2, l'input auditif (perception), les compétences en production orale et le rôle de la L1 (entre autres) soient les facteurs les plus importants pour expliquer

l'émergence de la structure sonore en L2, il est indéniable que les effets du visuel (et plus particulièrement les effets de l'orthographe sur la production orale) ont une incidence sur la prononciation de la L2.

# 3 Questions de recherche et méthodologie

Nous nous intéressons ici à étudier l'effet de la tâche (lecture *vs* imitation) et de l'orthographe en français L2 dans la production orale des hispanophones. Nous nous posons 2 types de questions : (i) Les erreurs de prononciation se distribuent-elles différemment en fonction de la tâche réalisée (imitation *vs* lecture) et/ou du niveau de langue des apprenants ? (ii) Quel est le rôle de l'orthographe dans l'émergence de ces erreurs ?

## 3.1 Protocole expérimental

Les données ont été collectées à partir du protocole IPFC-espagnol (Racine *et al.* 2012) auprès de 27 étudiants universitaires hispanophones du Mexique. Les participants ont réalisé plusieurs tâches : une tâche d'imitation sans support écrit, deux tâches de lecture de mots avec un support écrit, la lecture d'un texte, un entretien semi-guidé et une production en binômes (avec un autre apprenant). Pour la présente étude, nous analysons la production orale des 68 mots conçus dans le protocole IPFC-espagnol dans la tâche de lecture et d'imitation. Ces mots contenaient des phonèmes potentiellement difficiles pour les locuteurs hispanophones apprenant le français L2 (par exemple, /s/~/z/ dans les mots *hausse* vs *ose*, /u/ ~/y/ dans les mots *boule* vs *bulle*, l'articulation du /d/ en position initiale *vs* intervocalique dans les mots *dorer* vs *adorer*, etc.). Dans la tâche d'imitation, les participants ont écouté une série de 68 stimuli (mots isolés) et avaient 5 secondes entre chaque mot pour le reproduire. La tâche de lecture consistait à lire les mêmes mots que dans la tâche d'imitation mais devant un écran. Pour cette tâche, les participants devaient lire l'input affiché sur l'écran d'un ordinateur avec 3 secondes d'intervalle entre chaque mot.

## 3.2 Participants

Les participants étaient formés de 27 étudiants mexicains (16 femmes et 11 hommes) qui poursuivant leurs cours de formation universitaire à l'Université Nationale Autonome du Mexique. Ils poursuivaient, parallèlement à leurs cours universitaires, des cours de français L2 dans l'Ecole Nationale de Langues de cette université. Les étudiants ont été classés en deux groupes selon le niveau de leurs cours de français au moment de l'expérience : 9 étudiants intermédiaires ou A2(+) et 18 avancés ou B1(+). Leur âge moyen est de 23,9 ans (é.-t. 5,8). Tous les participants ont été enregistrés dans une pièce calme, dans les locaux de cette université.

## 3.3 Annotation

Le corpus analysé contient 3672 mots produits par les participants (68 mots x 2 tâches x 27 locuteurs). Deux phonéticiens ont encodé la production de tous les mots avec une transcription phonétique large (coarticulation exclue) de manière individuelle sous *Praat*. Ensuite, les annotateurs ont réalisé une correction manuelle avec la visualisation du spectrogramme. Enfin, dans une troisième étape, les annotations ont été confrontées afin de les harmoniser. En cas de divergences, les deux annotateurs se sont entendus sur une seule transcription.

### 3.4 Distances de Levenshtein

L'évaluation de la prononciation a été obtenue à partir des scores obtenus avec les distances de Levenshtein, mesures utilisées dans d'autres études en acquisition de L2 (cf. Weiling *et al.* 2014). Ces distances ont été calculées à partir de la transcription phonétique de la production des apprenants et la production attendue (transcription phonétique standard du français parisien). Cet algorithme permet de calculer le degré de différence entre la transcription de la cible et la prononciation réelle. Le score obtenu rend compte du nombre d'insertions, suppressions ou remplacements nécessaires pour transformer la chaîne de caractères de la transcription de la prononciation en L2 en la chaîne de la prononciation cible. L'algorithme permet également d'aligner automatiquement les segments identiques entre les deux annotations.

Un exemple est illustré dans la FIGURE 1 avec les valeurs obtenues avec les mots *hors* et *augmenter* dans la tâche de lecture. Dans ces exemples, les lignes rouges symbolisent les remplacements de phonèmes et les points verts les ajouts. Pour chacun de ces changements (erreurs de prononciation de l'apprenant), l'algorithme fournit la valeur de 1, tandis que pour les alignements sans changements (prononciation canonique), la valeur est de 0.

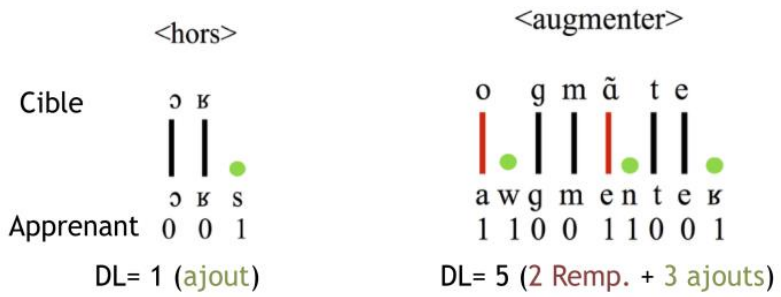



FIGURE 1: Exemple des valeurs obtenues via les distances de Levenshtein

Les distances ont été ensuite normalisées par rapport au nombre de phonèmes pour chaque mot afin d'éviter les effets de la variabilité du nombre de phones de chaque stimulus sur le nombre de changements nécessaires. Pour mesurer les effets de la tâche, nous avons calculé un pourcentage de similarité avec les distances de Levenshtein normalisées entre la production des apprenants et la prononciation normée.

Pour ce qui est des effets de l'orthographe, nous avons retenu l'analyse de la production de 4(archi)phonèmes en particulier, lesquels étaient associés à une graphie transparente/opaque : (i) /z/ écrit avec une graphie transparente <**z**oo> *vs* opaque <ba**s**e>, (ii) /O/ écrit avec une graphie transparente <p**o**rt> *vs* opaque <p**eau**>, (iii) /u/ écrit avec une graphie opaque <b**ou**le> et (iv) /y/ écrit également avec une graphie opaque <b**u**lle>. Il faut noter que l'opacité de ces deux derniers phonèmes est déterminée en fonction de la L1 des apprenants. En effet, les graphies <ou> et <u> correspondent aux sons [ou̯] et [u] respectivement en espagnol, mais aux sons [u] et [y] en français L2.

# 4 Résultats

Les données ont été analysées moyennant des modèles linéaires mixtes (Baayen 2008) avec le package *lme4* sous R (Bates *et al*. 2015). Les constantes (*intercepts*) aléatoires pour les locuteurs et les mots ont été évaluées dans ces modèles, ainsi que les pentes aléatoires par locuteur et par mot. La

puissance des variables prédictives a été estimée moyennant des tests de rapport de vraisemblance entre les modèles avec et sans les effets fixes.

## 4.1 Effets de la lecture et du niveau

La première question concernait la distribution des erreurs en fonction de la tâche et du niveau des apprenants. La FIGURE 2 montre que la lecture paraît affecter de manière négative la prononciation chez les participants. En revanche, ce n'est pas le cas pour les effets du niveau où nous pouvons observer que les pourcentages de similarité ne diffèrent pas entre les étudiants A2 et B1.

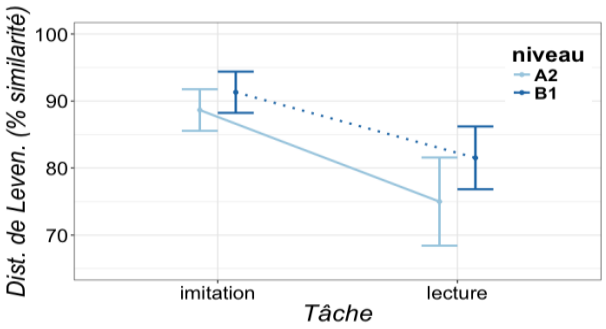



FIGURE 2: Effets de la tâche et du niveau dans la prononciation en français L2

Afin de confirmer ces observations, nous avons construit un modèle à effets mixtes (multinomial). Dans ce modèle, nous avons entré les valeurs des POURCENTAGES DE SIMILARITÉ selon les distances de Levenshtein comme variable dépendante, le NIVEAU des apprenants (A2 *vs* B1), la TACHE (imitation *vs* lecture) comme effets fixes, et les participants et les mots comme effets aléatoires. Les résultats montrent que le pourcentage moyen de similarité entre la prononciation réelle des étudiants et la prononciation normée est de 90,5% dans la tâche d'imitation, alors qu'en lecture, le degré de similarité descend à 79,9%. Les résultats statistiques ont confirmé que la TÂCHE a un effet sur la prononciation des apprenants ($\chi^2(2) = 28{,}17$, $p < 0{,}0001$). Autrement dit, la lecture fait augmenter les erreurs de prononciation de 10% selon nos métriques. Pour ce qui est du NIVEAU, les résultats montrent que les étudiants B1 arrivent à reproduire les mots avec la prononciation attendue avec 86,2% de similarité, tandis que les étudiants A2 le font avec un taux de similarité de 82,2%. Les résultats du modèle statistique montrent, en revanche, que cette différence n'est pas statistiquement significative ($p > 0{,}05$). Les résultats du modèle évaluant les effets d'une interaction TACHE*NIVEAU sur les POURCENTAGES DE SIMILARITÉ n'ont pas atteint le seuil de significativité non plus ($p > 0{,}05$). En d'autres mots, les effets négatifs de la TÂCHE (lecture) dans la prononciation sont similaires pour les deux niveaux de langue testés (cf. la similitude des pentes de la FIGURE 2).

Ces observations démontrent que les étudiants ne semblent pas améliorer leur prononciation selon le niveau. De fait, ces résultats montrent que les étudiants B1 sont autant affectés que les étudiants A2 par la tâche de lecture. Ces résultats confirment ce qui avait été soulevé par les études précédentes : les participants montrent une certaine sensibilité à la présence des graphies lors des tâches de lecture oralisée. En plus, cette influence de l'écrit sur l'oral ne semble pas diminuer avec l'apprentissage de la L2.

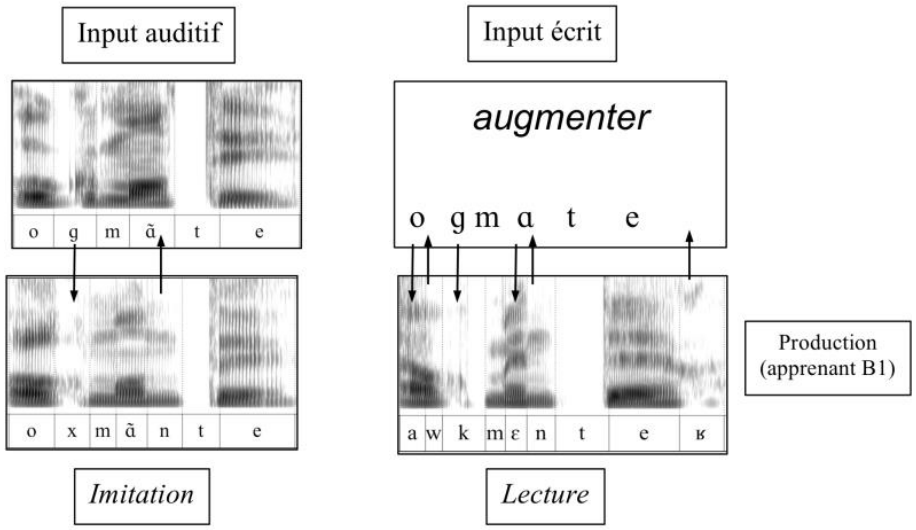



FIGURE 3: Ajouts et remplacements des phones pour le mot *augmenter* selon la tâche

La FIGURE 3 illustre les effets de la tâche dans nos données. Nous pouvons constater que le nombre de remplacements (flèches vers le bas) et ajouts (flèches vers le haut) s'élèvent dans la tâche de lecture par rapport à la tâche d'imitation. Cette analyse montre que la présence de l'input écrit modifie négativement la prononciation chez les participants. Dans cet exemple, il est intéressant de noter que le nombre d'insertions de phones déclenchés par la lecture est très probablement dû à une association erronée de la production oralisée des graphies silencieuses dans l'input écrit. Cette analyse réside dans le fait que l'insertion et le remplacement des phones de cet exemple ne peuvent pas être directement attribuables à l'influence de la phonétique/phonologie de la L1. A tout le moins, le remplacement de la voyelle [o] par la suite [aw], et l'ajout du [ʁ] final dans le mot <**au**gmente**r**> dans la tâche de lecture peut être difficilement attribuable à l'influence de la phonologie/phonétique de la L1 de l'apprenant, mais plutôt à une mauvaise association des graphies avec les sons.

## 4.2 Effets de l'orthographe

Le deuxième objectif de cette étude était d'examiner les effets de l'orthographe à partir de la réalisation des phonèmes associés à des graphies transparentes *vs* opaques. Notre hypothèse était que la présence d'une graphie opaque devrait déclencher des erreurs de prononciation, à la différence de la présence d'une graphie transparente, laquelle devrait favoriser la prononciation. attendue. Si c'était le cas, nous pourrions donner des arguments pour montrer que les erreurs des apprenants (remplacement d'un phonème par un autre) sont induites exclusivement par la présence de la graphie. Ainsi, nous avons évalué si la prononciation du phonème /z/ émergeait comme [z] dans les mots où un tel phonème était associé à une graphie transparente (<**z**oo>) *vs* opaque (<ba**s**e>) dans la tâche de lecture (5 mots pour chaque cas). La même hypothèse était formulée pour l'articulation de l'archiphonème /O/ : nous avons comparé les mots où cet archiphonème était associé à une graphie transparente (/ɔ/ pour <p**o**rt>) *vs* opaque (/o/ pour <p**eau**>) par rapport à la L1 de l'apprenant (4 mots différents pour chaque cas).

Les phonèmes français /u/ *vs* /y/ seraient tous les deux associés à des graphies opaques pour l'apprenant selon les règles de leur L1 (*v.* supra). Nous avons donc comparé si les effets de l'orthographe étaient les mêmes pour chacun de ces cas (<b**ou**le> vs <b**u**ll>, 3 mots différents pour chaque phonème.

La FIGURE 4 montre les pourcentages moyens de réussite entre le phonème produit par les participants et la cible attendue en fonction des graphies transparentes (bleu clair) et opaques (bleu foncé). Cette figure montre que les graphies opaques ont des effets négatifs dans la prononciation comme il était attendu, car les taux de succès descendent dans les deux cas.

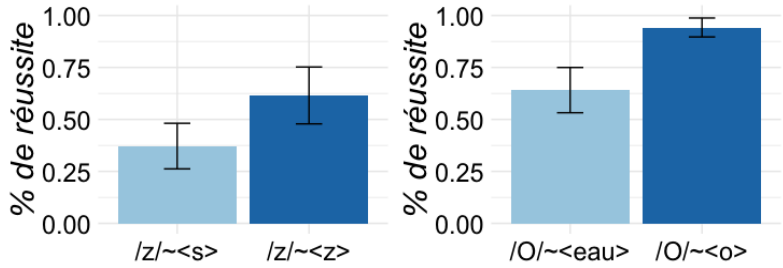



FIGURE 4: Effets des graphies opaques *vs* transparentes dans la prononciation en français L2

Nous avons construit différents modèles à effets mixtes (binomiaux) afin de corroborer ces observations où nous avons entré POURCENTAGES DE RÉUSSITE (phonème correct *vs* incorrect) comme variable dépendante, NIVEAU (A2 *vs* B1) et GRAPHIES (transparente *vs* opaque) comme variables fixes, et participants et mots comme effets aléatoires. Les résultats statistiques montrent un effet des GRAPHIES : les erreurs de prononciation augmentent avec la lecture des graphies opaques par rapport aux graphies transparentes ($\chi^2(2) = 8{,}87$, $p < 0{,}01$). Ainsi, le phonème /z/ est articulé comme [z] dans 61% des cas avec la graphie transparente <z> contre 37% avec la graphie opaque <s>, laquelle a été prononcée massivement comme [s]. Les effets du NIVEAU ne sont pas statistiquement significatifs ($p > 0{,}05$), cela montrant que les deux groupes ont des taux de réussite similaires (49% pour A2, et 50% pour B1). Les résultats montrent que les effets de l'interaction NIVEAU*GRAPHIE n'atteignent pas non plus le seuil de significativité ($p > 0{,}05$). En d'autres termes, les effets des GRAPHIES sur la prononciation sont similaires dans les deux groupes d'apprenants.

Pour le cas de l'archiphonème /O/, il est articulé, soit comme [o] ou soit comme [ɔ], dans 94% des cas lorsque ce phonème est associé à la graphie <o> contre 64% des cas avec les graphies <eau>, contexte où les apprenants ont produit le son [ø] dans la plupart des cas. Le test statistique évaluant les effets du facteur GRAPHIES montre que les graphies opaques nuisent à la prononciation de cet archiphonème ($\chi^2(2) = 35{,}83$, $p < 0{,}0001$). Pour ce qui est des effets du NIVEAU, le taux de réussite est de 73% pour le groupe A2 et de 84% pour le groupe B1. Toutefois, ces différences ne sont pas statistiquement significatives ($p > 0.05$). Les résultats du modèle statistique évaluant l'interaction NIVEAU*GRAPHIES n'atteignent pas le seuil de significativité ($p > 0{,}05$), cela montrant que les effets négatifs des graphies opaques affectent de manière similaire la prononciation de cet archiphonème dans les deux groupes d'apprenants.

Enfin, nous avons analysé dans quelle mesure les deux types de graphies opaques selon la L1 des apprenants avaient une incidence dans la prononciation et avons examiné si l'une d'entre elles s'avéraient comme plus problématique pour les apprenants. C'est le cas de la production du phonème /u/ et /y/ associés aux graphèmes <ou> et <u> respectivement. Les résultats montrent un effet des GRAPHIES : la graphie <u> fait augmenter les erreurs de prononciation par rapport aux graphies <ou> ($\chi^2(2) = 8{,}55$, $p < 0{,}01$). La production du phonème /u/ émerge comme [u] dans 88% des cas alors que le phonème /y/ émerge comme [y] dans 51% des cas. Ceci montre que le graphème <ou> trouble moins les participants que le graphème <u> où ce dernier était massivement articulé comme [u]. Ceci contraste avec les résultats de la tâche d'imitation, où le phonème /y/ émerge comme [y] dans le 96%

des cas. Les résultats montrent aussi que le facteur NIVEAU n'a pas d'effets dans la prononciation ($p > 0,05$). Les valeurs de $p > 0,05$ du test statique évaluant les effets de l'interaction NIVEAU*GRAPHIES montre que, comme dans les deux cas précédents, les effets négatifs de ces deux graphies sur la prononciation sont similaires dans les deux groupes d'apprenants.

# 5 Discussion & Conclusion

Les résultats reportés dans les sections précédentes permettent de faire plusieurs constatations. D'une part, la tâche de lecture a des effets négatifs sur la prononciation des apprenants : elle peut déclencher une mauvaise association des graphies-sons, ou bien l'insertion de phonèmes due à la présence de graphies silencieuses. D'autre part, nos résultats montrent que le niveau de compétence phonique dans les deux groupes est relativement homogène. Cela est d'autant plus surprenant que la différence de niveau de langue est relativement importante : A2(+) *vs* B1(+). Ces observations suggèrent que les compétences de prononciation des apprenants hispanophones selon leur niveau de langue semblent être immuables. De fait, nos résultats montrent que l'influence négative de la tâche de lecture ne diminue pas avec l'apprentissage de la L2.

Ensuite, nous avons trouvé que les graphies opaques déclenchent des erreurs de prononciation pour les quatre phonèmes étudiés ici. Le fait que le remplacement des phones augmente avec les graphies opaques en rapport avec les graphies transparentes confirme qu'il existe des rapports entre la compétence phonétique et l'orthographe. Nous montrons que, dans tous les cas, le niveau des apprenants n'a aucune incidence dans la prononciation : les effets négatifs de l'orthographe affectent de manière similaire les deux niveaux testés. Ainsi, il semblerait que les apprenants B1 ne semblent pas surmonter les interférences de l'orthographe par rapport aux étudiants A2.

Finalement, si les résultats présentés ici n'abordent pas tous les phonèmes en français, ni toutes leurs possibilités de représentation orthographique, ils confirment que la compétence phonique n'est pas dissociée des compétences visuelles, en l'occurrence, l'association graphies-sons. Ces résultats permettent d'avancer la question suivante : le transfert phonologique négatif de la L1 est-il renforcé par la présence de l'écrit ? L'un des objectifs de notre future recherche est de mieux examiner les traits phonologiques concernés dans le transfert de la L1-L2 attribuables à la phonologie et à l'orthographe. Cela permettrait d'avoir des explications plus claires sur les effets de l'orthographe dans le développement de la compétence phonétique en français L2. Il n'en reste pas moins que nous souhaitons souligner les effets positifs de la tâche d'imitation : la perception auditive semble être en définitive une technique qui n'entraînerait que des bénéfices dans l'apprentissage de la structure sonore de la L2.

# 6 Références

BAAYEN, R.H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge : Cambridge University Press.

BASSETTI, B. (2017). Orthography affects second language research: double letters and geminate production in English. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 43(11), 1835-1842.

BATES, D., MAECHLER, M., BOLKER, B. & WALKER, S. (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software,* 67(1), 1-48.

BEST, C.T. (1995). A direct realist view of cross-language speech perception. In W. STRANGE (éd.), *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research.* Timonium, MD: York Press, 171-204.

CATACH, N. (1991). L'orthographe. Paris: PUF, 9e édition.

CHEVROT, J.-P. & MALDEREZ, I. (1999). L'effet Buben : de la linguistique diachronique à l'approche cognitive (et retour). *Langue française,* 124, 104-125.

DETEY, J., DURAND J. & NESPOULOUS J.-L. (2005). Interphonologie et représentations orthographiques. Le cas des catégories /b/ et /v/ chez des apprenants japonais de Français Langue Etrangère. *Revue Parole*, 34-36, 140-185.

ESCUDERO, P. (2005). *Linguistic Perception and Second Language Acquisition: Explaining the attainment of Optimal Phonological Categorization.* Thèese de Doctorat. Université d'Utrecht.

FLEGE, J.E. (1995). Second language speech learning: Theory, findings and problems. In W. STRANGE (éd.), *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research.* Timonium, MD: York Press, 233-277.

FRAUENFELDER, U.H., SEGUI, L. & DIJKSTRA, T. (1990). Lexical effects in phonemic processing: facilitatory or inhibitory? *Journal of Experimental Psychology: Human Perception and Performance,* 16 (1), 77-91.

RACINE, I., DETEY, S., ZAY, F. & KAWAGUCHI, Y. (2012). Des atouts d'un corpus multitâches pour l'étude de la phonologie en L2: l'exemple du projet "Interphonologie du français Contemporain" (IPFC). In KAMBER, A., SKUPIENS, C. (éd.), *Recherches récentes en FLE.* Berne: Peter Lang, 1-19.

SILVEIRA, R. (2007). The role of task-type and orthography in the production of word-final consonants. *Revista de Estudos da Linguagem,* 15(1), 143-176.

YOUNG-SCHOLTEN, M. (2002). Orthographic input in L2 phonological development. In. P. BURMEISTER & A. RHODE (éd.), *An integrated view of language development: Papers in honour of Henning Wode*. Trier, Allemagne: Verlag, 263-279.

WIEILING, M. et al. (2014). Measuring foreign accent strength in English : Validating Levenshtein Distance as a Measure. *Language Dynamics and Change*, 4(2), 253-269.

ZAMPINI, M. (1994). The role of native language and tasks formality in the acquisition of Spanish Spirantization. *Hispania*, 77(3), 470-481.

XXXIIe Journées d'Études sur la Parole
4-8 juin 2018, Aix-en-Provence, France

# Etude de performance des réseaux neuronaux récurrents dans le cadre de la campagne d'évaluation Multi-Genre Broadcast challenge 3 (MGB3)

Salima Mdhaffar Antoine Laurent Yannick Estève
Laboratoire d'Informatique de l'Université du Mans (LIUM), Avenue Laennec, Le Mans, France
prenom.nom@univ-lemans.fr

RÉSUMÉ

Ces dernières années, l'utilisation des réseaux neuronaux est devenue incontournable dans de nombreux domaines et notamment en traitement automatique des langues. Le travail présenté dans cet article s'inscrit dans le cadre de leur utilisation dans le domaine de la reconnaissance automatique de la parole. Nous présentons les résultats obtenus par des réseaux neuronaux récurrents (RNN) de natures différentes (LSTM, GRU, GRU-Highway) sur les données de la campagne d'évaluation MGB 3. Les données de cette campagne, qui n'est pas encore terminée, correspondent à des enregistrements d'émissions très diverses de la chaîne de télévision britannique BBC. Nos expériences offrent une comparaison des résultats des différents RNN et comment, en combinant des réseaux de neurones récurrents et des modèles de langage N-gram classiques modélisant les phrases dans les deux sens de lecture, il est possible d'améliorer de manière très significative les performances d'un système de reconnaissance de la parole.

ABSTRACT

**Studying performances of recurrent neural networks in the context of the Multi-Genre Broadcast challenge 3 (MGB3) evaluation campaign**

In recent years, the use of neural networks has become indispensable in many fields related to automatic language processing. The work presented in this paper explores the use of several recurrent neural network (RNN) variants (LSTM, GRU and GRU-Highway) in automatic speech recognition in the context of MGB 3 evaluation campaign. The data for this campaign, which is still undergoing, corresponds to a wide variety of emissions recorded from the British television channel BBC. Besides comparing the impact of different RNN, we also show how a combination of RNN and N-gram language models processing the sentences in both reading directions, significantly improves the performance of a speech recognition system.

MOTS-CLÉS : Reconnaissance Automatique de la Parole, Modèle de Langage, Réseaux de Neurones Récurrents, Modèle de Langage à l'Arrière, Interpolation.

KEYWORDS: Automatic Speech Recognition, Language Model, Recurrent Neural Networks, Reverse Language Model, Interpolation.



10.21437/JEP.2018-20

# 1 Introduction

De nos jours, la reconnaissance automatique de la parole est un domaine très actif et a connu une évolution technologique et scientifique très rapide. La raison primordiale de cette évolution est l'utilisation des réseaux de neurones dans la modélisation acoustique et linguistique.

Un modèle de langage (ML) constitue un composant très important dans un système de reconnaissance de la parole. Un ML a pour but de guider le décodeur à choisir la séquence de mots la plus probable. Depuis longtemps, les modèles n-gram sont les plus employés dans la modélisation statistique des langues du fait de leur mise en œuvre rapide et de leur robustesse. Un modèle de langage n-gram estime la probabilité d'apparition d'un mot sachant les n-1 mots qui le précèdent $P(m_i|m_{i-n+1}, ..., m_{i-2}, m_{i-1})$.

Cependant, les modèles N-gram présentent certaines limites. Ils modélisent mal les contraintes à longue distance et nécessitent l'utilisation de techniques de lissage (Chen & Goodman, 1996) pour pallier le problème des événements non vus dans le corpus d'apprentissage. Les techniques les plus connues, sont les techniques de décomptes de Good-Turing (Good, 1953), de Witten-Bell (Witten & Bell, 1991) et de Kneser-Ney (Kneser & Ney, 1995) qui utilisent toute une stratégie de repli (*back-off*).

Malgré la multitude des méthodes de lissage développées, les modèles n-gram prennent mal en compte le fait que les mots dont le sens ou la morphosyntaxe sont proches peuvent s'apparaître dans des contextes similaires. Ceci est dû à la représentation des mots dans un espace discret (le vocabulaire) où il n'existe aucun partage d'information morphologique, syntaxique ou sémantique entre les mots.

Les modèles n-gram basés sur les classes (Brown *et al.*, 1992) ont été introduits afin d'aborder ce problème en regroupant les mots et les contextes dans des classes en fonction de leurs utilisations. L'exploitation des informations relatives aux classes permet d'améliorer la généralisation d'estimation des modèles de langage. Cependant, les problèmes auxquels sont confrontés ces modèles de type n-classes sont nombreux. Le premier problème est que ce type de modèle exige d'avoir un corpus d'apprentissage pré-étiqueté. L'étiquetage manuel est une tâche très coûteuse. En plus, ces modèles possèdent aussi le même inconvénient que celui des modèles N-grams en ce qui concerne la taille de l'historique pris en compte. En effet, la majorité des modèles présentés dans la littérature se limitent à un historique restreint à 3 ou 4 classes (ou mots).

Ces dernières décennies, l'utilisation des réseaux de neurones dans la modélisation du langage a connu beaucoup de succès et a permis l'obtention de performances très intéressantes. Le principe de base de ces modèles s'appuie sur la projection des mots du contexte dans un espace continu ce qui permet d'exploiter la notion de similarité entre les mots. La force des modèles de langage neuronaux réside dans leur capacité de bien généraliser les N-grams non vus car les mots similaires vont avoir probablement le même contexte. Plusieurs travaux dans la littérature prouvent que les modèles de langage neuronaux donnent des résultats plus performants que les modèles de langage n-gram (Mikolov *et al.*, 2011; Schwenk & Gauvain, 2002; Schwenk, 2007).

Dans cet article, nous étudions l'utilisation des modèles de langage neuronaux récurrents dans le cadre de la tâche de reconnaissance automatique de la parole. Nous explorons également l'interpolation de plusieurs types de modèles récurrents et de modèles n-gram modélisant les phrases dans les deux sens de lecture. Ces travaux ont été réalisé dans le cadre de la participation du LIUM dans la campagne d'évaluation MGB.

La suite de l'article est organisée comme suit. Les réseaux de neurones récurrents sont détaillés dans la section 2. Dans la section 3, nous décrivons en détail les implémentations effectuées. La section 4 est consacrée à la présentation du corpus sur lequel nous évaluons les modèles, et présente les différents résultats obtenus. La section 5 conclut l'article.

# 2 Etat de l'art sur les réseaux de neurones récurrents (RNN)

Plusieurs variantes de réseaux de neurones récurrents existent dans la littérature. Les modèles de langage basés sur les réseaux de neurones récurrents sont composés d'au minimum trois couches, à savoir une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie. L'entrée du réseau à l'instant $t$ est $x(t)$, la sortie est notée $y(t)$, et $h(t)$ est la couche cachée. La couche d'entrée (équation 1) est formée d'un vecteur $w(t)$ qui contient la représentation continue du mot actuel et d'un vecteur $s(t-1)$ qui représente les valeurs de sortie dans la couche cachée à partir de l'étape précédente.

$$x(t) = w(t) + s(t-1). \tag{1}$$

Les vecteurs $w(t)$ et $s(t-1)$ sont concaténés dans un seul vecteur afin de former l'entrée de la couche cachée $s(t)$. La fonction d'activation $g$ (équation 2) de la couche cachée est une fonction non linéaire (généralement une sigmoïde (Han & Moraga, 1995)).

$$h_t = g(w(t), s(t-1)) \tag{2}$$

La couche de sortie $y_t$ (équation 3) est constituée d'un nombre de neurones qui est égal à la taille du vocabulaire $v$ ou à la taille de la *shortlist*[1]. Le but de la couche de sortie est de fournir les probabilités de chaque mot w dans le vocabulaire ou de la *shortlist* en fonction de l'historique. La couche de sortie utilise la fonction d'activation softmax (Gao & Pavel, 2017) afin de garantir que la somme des probabilités est égale à 1.

$$y_t = softmax(v * h_t + b) \tag{3}$$

où $b$ est un vecteur de biais.

Bien que les RNN puissent, en théorie, modéliser des dépendances infiniment longues, ils ne sont pas capables de mémoriser des historiques de grande taille. Les réseaux de neurones utilisant des unités de type *Long Short-Term Memory* (LSTM) sont des variantes des réseaux de neurones récurrents dont les unités de base intégrent différentes portes (Hochreiter & Schmidhuber, 1997) (en anglais *gate*), permettant d'écrire, de mettre à jour ou de lire une mémoire contextuelle, à partir d'informations vues précédemment. Ces portes permettent aux LSTMs de modéliser plus efficacement les dépendances longue distance.

Un LSTM est composé d'une mémoire et de trois portes. La porte d'oubli $f$ (forget) contrôle quelle est la partie de la cellule précédente qui sera oubliée. La porte d'entrée $i$ (input) doit choisir les informations pertinentes qui seront transmises à la mémoire. La sortie $o$ (output) contrôle quelle partie de l'état de la cellule sera exposée en tant qu'état caché.

1. la *shortlist* est généralement une liste contenant les mots les plus fréquents dans le corpus d'apprentissage (Schwenk, 2007)

Une autre variation des LSTMs sont les *Gated Recurrent Unit* (GRU) (Cho *et al.*, 2014) qui sont plus simples que le LSTM. Ils ont l'avantage d'être moins coûteux en calculs car ils possèdent moins de paramètres. Ils incorporent seulement deux types de portes au lieu de trois : une porte de réinitialisation $r$ (reset) qui détermine comment combiner la nouvelle entrée avec la mémoire précédente et une porte de modification $u$ (update) qui permet de décider si l'état caché $h$ doit être mis à jour avec le nouvel état caché $h$ ou non.

Une autre extension du GRU est le GRU-Highway. Le réseau highway (Srivastava *et al.*, 2015) est une approche proposée pour optimiser les réseaux et augmenter leurs profondeurs. Le GRU-Highway sert à calculer une sortie qui est une combinaison entre l'entrée et le résultat du GRU. On a $h_t^{(gru)}$ le résultat du GRU classique, $x_t$ est l'entrée à l'instant t et $g$ est une fonction sigmoïde. Les équations pour le GRU-Highway sont les suivantes :

$$g_t = g(x(t), s(t-1)) \tag{4}$$

$$h_t = g_t.h_t^{(gru)} + (1 - g_t).x_t \tag{5}$$

Dans cette étude, nous nous intéressons aux réseaux de neurones récurrents dans le cadre de la modélisation de langage pour la reconnaissance de la parole (RAP).

# 3 Implémentation

Afin de comparer les performances de différents RNNs, plusieurs implémentations ont été réalisées. Cette section détaille ces implémentations.

## 3.1 Implémentation d'un modèle de langue N-gram

Un modèle de langage 3-gram a été construit comme système de base afin de pouvoir mesurer l'amélioration en terme de taux d'erreur de mots (**W**ord **E**rror **R**ate : WER) apportée par les différents modèles neuronaux. Ainsi, ce modèle trigram va être utilisé pour le décodage de la parole afin de construire une liste de N meilleures hypothèses (la liste de N-best). Ce modèle trigram va servir aussi par la suite pour effectuer une interpolation linéaire avec les modèles de langage neuronaux. Par interpolation d'un modèle de langage n-gram avec un modèle de langage RNN, nous entendons une combinaison linéaire de probabilités des phrases (N-best) obtenues à partir de différents modèles, avec des coefficients de pondération pour chaque modèle.

Pour l'interpolation, nous avons aussi construit un modèle 4-gram pour le comparer avec un modèle 3-gram.

## 3.2 Implémentation de trois modèles RNN : GRU, LSTM et GRU-Highway

Trois modèles récurrents : GRU, LSTM et GRU-Highway sont implémentés en utilisant l'outil CUED-RNNLM [2] (Chen *et al.*, 2016). CUED-RNNLM est un outil libre destiné à la modélisation

2. http ://mi.eng.cam.ac.uk/projects/cued-rnnlm/

du langage avec les réseaux de neurones. Il comprend plusieurs types de réseaux de neurones tels que le modèle de langage neuronal "Feed forward" et plusieurs variétés de RNN. La boîte à outils CUED-RNNLM fournit aussi des recettes pour diverses fonctions, notamment l'évaluation de la perplexité, le ré-évaluation de N-best, etc.

Pour garantir une comparaison équitable entre tous les modèles, nous avons utilisé les mêmes réglages et paramètres pour tous les modèles RNN. Le nombre de couches cachées est 2. La taille de la couche cachée est de 200. La taille du vocabulaire de sortie est de 30000 mots. Un modèle avec 60000 mots a également été implémenté pour évaluer l'apport de l'augmentation de la taille du vocabulaire.

### 3.3 Implémentation d'un RNN arrière

Généralement, un modèle de langage est entraîné dans une seule direction : du passé au futur. Cependant, même les données du futur peuvent aider un modèle à estimer la probabilité d'apparition d'un mot. Alors, il est avantageux de construire un modèle de langage dans lequel l'ordre des mots est inversé. Un modèle de langage dont l'ordre des mots est inversé estime la probabilité d'un mot sachant le contexte futur $P(w_\alpha|w_{\alpha+1}, ..., w_{\alpha+n})$ où $\alpha$ est l'indice du mot courant et n le nombre de mots dans le contexte. Le RNN arrière que nous avons implémenté, est similaire à un RNN avant à l'exception que pendant l'apprentissage du modèle la phrase est donnée à l'envers : le premier mot devient le dernier et le dernier mot devient le premier.

Le but d'implémenter un RNN arrière est d'être capable de prendre en compte l'information passée et future mémorisée par interpolation des deux modèles arrière et avant pour effectuer les estimations.

Par interpolation d'un modèle de langage RNN arrière avec un modèle de langage RNN avant, nous entendons une combinaison linéaire de probabilités des phrases (N-best) obtenues à partir du modèle RNN avant et à partir du modèle RNN arrière.

Un modèle 3-gram et un modèle 4-gram arrières ont également été entrainés selon le même principe dans le but d'interpoler les quatre modèles : modèle N-gram avant, modèle N-gram arrière, modèle RNN avant, modèle RNN arrière.

## 4 Résultats expérimentaux

### 4.1 Description du système de RAP

Le système de RAP utilisé pour les expériences présentées dans ce papier est un système préliminaire développé dans le cadre de la campagne d'évaluation MGB 2017. Un premier système de type HMM-GMM contexte dépendant (3-phones), MFCC+LDA+MLLT a été entrainé pour générer les alignements du corpus. Une technique de perturbation de la vitesse (Ko *et al.*, 2015) a été utilisée pour multiplier par 3 la quantité des données d'entrainement.

Ensuite, un modèle de type chain-TDNN (Lattice-free MMI TDNN (Povey *et al.*, 2016)) avec un entraiement discriminant visant à minimiser le risque bayésien sur les états (sMBR - (Kingsbury *et al.*, 2012)) a été entrainé. Des iVecteurs ont également été mis en oeuvre pour l'adaptation instantanée des réseaux de neurones (Saon *et al.*, 2013).

Les phonétisations des mots sont obtenues à l'aide d'un lexique réalisé manuellement, dérivé de Combilex, fourni par les organisateurs de la campagne d'évaluation.

## 4.2 Corpus

Afin de valider notre implémentation du réseau de neurones et comparer nos résultats, nous avons mené des expériences sur les données de la compagne d'évaluation MGB 2017 [3] (**M**ulti **G**enre **B**roadcast) pour la tâche en anglais. MGB est une compagne d'évaluation pour la transcription automatique des émissions TV.

Les données de la compagne d'évaluation MGB 2017 comprennent environ 328 heures d'audio enregistrées sur sept semaines à partir de toutes les chaînes de télévision de BBC (BBC1, BBC2, BBC3, BBC4, etc). Les données couvrent une grande variété de genres (documentaires, actualités, drames ,etc). Quelques statistiques du corpus sont présentées dans le tableau 1.

La campagne MGB n'étant pas encore terminée, les données de test ne sont pas encore fournies. Par conséquent, nous avons éliminé les données de développement du corpus d'apprentissage et nous les avons utilisées pour évaluer notre système.

| Corpus | Corpus d'apprentissage | Corpus de développement |
|---|---|---|
| # segments | 237068 | 5856 |
| # locuteurs | 2719 | 302 |
| Durée | 324h | 4h37 |

TABLE 1 – Statistiques du corpus MGB3

Ainsi, pour l'apprentissage des modèles de langage, nous avons utilisé les données fournies par la compagne d'évaluation MGB 2017. En effet, la compagne d'évaluation MGB fournit des données pour la modélisation acoustique ainsi que des données pour la modélisation linguistique [4]. Quelques statistiques des données d'apprentissage des modèles de langage sont présentées dans le tableau 2.

| | Corpus d'apprentissage |
|---|---|
| # mots total | 645758382 |
| Vocabulaire | 757748 |
| Vocabulaire utilisé pour les ML n-gram | 164000 |

TABLE 2 – Statistiques du corpus d'apprentissage pour la modélisation linguistique

## 4.3 Analyse des résultats

La qualité des modèles de langage implémentés est évaluée en terme de gain en WER. WER (Pallett, 2003) est la métrique d'évaluation couramment utilisée dans la littérature pour l'analyse des performances d'un système de reconnaissance de la parole. Elle se calcule comme suit :

$$WER = \frac{S + I + D}{N} \tag{6}$$

3. http ://www.mgb-challenge.org/
4. http ://www.mgb-challenge.org/download.html

Où S est le nombre de mots remplacés par le système, I est le nombre de mots insérés par le système, D est le nombre de mots supprimés par le système et N est le nombre total de mots dans une phrase.

Le tableau 3 présente les différents résultats expérimentaux. Afin d'obtenir ces résultats, deux passes sont effectués : la première passe consiste à obtenir une liste de N-best (les N meilleures hypothèses) avec N=200 en utilisant le système de reconnaissance décrit, la deuxième passe consiste à attribuer des scores à ces N-bests en utilisant les modèles de langage implémentés. La troisième colonne présente les résultats en terme de WER. La colonne 4 ($\delta$) présente le gain absolu par rapport au système de base sans effectuer la deuxième passe (1-best). Les poids d'interpolation utilisés sont 0,5 dans le cas d'interpolation de deux modèles et 0,25 dans le cas d'interpolation de quatre modèles.

| | | WER % | $\delta$ |
|---|---|---|---|
| 1 | 3-gram (système de base) | 24,7 | - |
| 2 | LSTM | 22,6 | 2,1 |
| 3 | GRU | 22,5 | 2,2 |
| 4 | GRU-Highway | 22,3 | 2,4 |
| 5 | LSTM + 3-gram | 22,1 | 2,6 |
| 6 | GRU + 3-gram | 22,0 | 2,7 |
| 7 | GRU + 4-gram | 21,7 | 3 |
| 8 | GRU-Highway + 3-gram | 21,6 | 3,1 |
| 9 | GRU + 3-gram (\|shortlist\|= 60K) | 21,8 | 2,9 |
| 10 | GRU + 4-gram (\|shortlist\|= 60K) | 21,6 | 3,1 |
| 11 | GRU arrière | 22,5 | 2,2 |
| 12 | GRU arrière + 3-gram arrière | 22,2 | 2,5 |
| 13 | GRU arrière + 3-gram arrière + GRU avant + 3-gram avant | 21.6 | 3,1 |
| 14 | GRU arrière + 4-gram arrière + GRU avant + 4-gram avant | 21.4 | 3,3 |

TABLE 3 – Résultats obtenus

Les résultats obtenus par les différents types de réseaux de neurones sont meilleurs que ceux obtenus avec le système de base (1) : decodé avec un modèle n-gram sans passer par la deuxième passe du rescoring.

Les résultats expérimentaux présentés dans le tableau 3 montrent que le GRU-Highway (4) a donné un résultat plus performant que le LSTM (2) et le GRU (3) en terme de WER (23,3% WER pour le GRU-Highway, 22,6 %WER pour le LSTM, 22,5 %WER pour le GRU).

L'interpolation de RNN avec un modèle n-gram permet une réduction 0,5% absolu de WER dans le cas de GRU+3-gram (6) par rapport au GRU (3).

En plus, l'interpolation avec un modèle 4-gram est plus performante que l'interpolation avec un modèle 3-gram. Le résultat d'interpolation d'un modèle GRU avec un 4-gram (7) est 21,7% WER par contre le résultat de l'interpolation d'un modèle GRU avec un 3-gram (6) est 22% WER.

Les résultats obtenus montrent que notre proposition de combiner un RNN arrière avec un RNN avant est utile et améliore significativement les résultats en termes de WER (13,14). Ceci confirme l'utilité de l'utilisation des informations du contexte passé et futur à la fois.

Enfin, nous avons montré que l'utilisation de 60000 mots dans la shortlist (9,10) a donné de meilleurs résultats par rapport à l'utilisation de 30000 mots dans shortlist (6,7).

# 5 Conclusion

Le travail présenté dans cet article est une étude portant sur l'utilisation des réseaux neuronaux pour la modélisation du langage dans le cadre de la reconnaissance automatique de la parole. Nous nous sommes intéressés en particulier aux réseaux neuronaux récurrents. Plusieurs types de réseaux neuronaux récurrents ont été évalués et expérimentés. Ainsi, nous avons exploré dans ce travail l'interpolation des modèles de langage neuronaux avec des modèles de langage n-gram ainsi que l'interpolation avec un modèle de langage récurrent arrière. Les expériences sont effectuées dans le cadre de la campagne d'évaluation MGB 2017. Les résultats obtenus montrent que les réseaux de neurones récurrents ainsi que les deux types d'interpolations apportent des améliorations significatives en terme de taux d'erreur mots.

# Remerciements

Nous remercions l'agence ANR pour son financement à travers le projet PASTEL sous le numéro de contrat ANR-16-CE33-0007.

# Références

BROWN P. F., DESOUZA P. V., MERCER R. L., PIETRA V. J. D. & LAI J. C. (1992). Class-based n-gram models of natural language. *Computational linguistics*, **18**(4), 467–479.

CHEN S. F. & GOODMAN J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, p. 310–318 : Association for Computational Linguistics.

CHEN X., LIU X., QIAN Y., GALES M. & WOODLAND P. C. (2016). Cued-rnnlm—an open-source toolkit for efficient training and evaluation of recurrent neural network language models. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, p. 6000–6004 : IEEE.

CHO K., VAN MERRIËNBOER B., GULCEHRE C., BAHDANAU D., BOUGARES F., SCHWENK H. & BENGIO Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv :1406.1078*.

GAO B. & PAVEL L. (2017). On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv :1704.00805*.

GOOD I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**(3-4), 237–264.

HAN J. & MORAGA C. (1995). The influence of the sigmoid function parameters on the speed of backpropagation learning. *From Natural to Artificial Neural Computation*, p. 195–201.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long short-term memory. *Neural computation*, **9**(8), 1735–1780.

KINGSBURY B., SAINATH T. N. & SOLTAU H. (2012). Scalable minimum bayes risk training of deep neural network acoustic models using distributed hessian-free optimization. In *Thirteenth Annual Conference of the International Speech Communication Association*.

KNESER R. & NEY H. (1995). Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, p. 181–184 : IEEE.

KO T., PEDDINTI V., POVEY D. & KHUDANPUR S. (2015). Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.

MIKOLOV T., KOMBRINK S., BURGET L., ČERNOCKỲ J. & KHUDANPUR S. (2011). Extensions of recurrent neural network language model. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, p. 5528–5531 : IEEE.

PALLETT D. S. (2003). A look at nist's benchmark asr tests : past, present, and future. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, p. 483–488 : IEEE.

POVEY D., PEDDINTI V., GALVEZ D., GHAHREMANI P., MANOHAR V., NA X., WANG Y. & KHUDANPUR S. (2016). Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Interspeech*, p. 2751–2755.

SAON G., SOLTAU H., NAHAMOO D. & PICHENY M. (2013). Speaker adaptation of neural network acoustic models using i-vectors. In *ASRU*, p. 55–59.

SCHWENK H. (2007). Continuous space language models. *Computer Speech & Language*, **21**(3), 492–518.

SCHWENK H. & GAUVAIN J.-L. (2002). Connectionist language modeling for large vocabulary continuous speech recognition. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 1, p. I–765 : IEEE.

SRIVASTAVA R. K., GREFF K. & SCHMIDHUBER J. (2015). Highway networks. *arXiv preprint arXiv :1505.00387*.

WITTEN I. H. & BELL T. C. (1991). The zero-frequency problem : Estimating the probabilities of novel events in adaptive text compression. *Ieee transactions on information theory*, **37**(4), 1085–1094.

# Étude des variations de fréquence fondamentale relatives au genre chez des bilingues Anglais/Français

Erwan Pépiot[1], Aron Arnold[2]
(1) TransCrit (groupe LECSeL), 2 rue de la liberté, 93526 Saint-Denis, France
(2) VALIBEL – Université Catholique de Louvain, Place Blaise Pascal 1, 1348 Louvain-la-Neuve, Belgique
erwan.pepiot@free.fr, aron.arnold@uclouvain.be

## RESUME

La présente étude porte sur les productions de locutrices et locuteurs bilingues anglais/français lors d'une tâche de lecture de phrases et en parole semi-spontanée. La fréquence fondamentale (F0) moyenne, la plage de variation de F0 et l'écart-type de F0 ont été mesurés dans les deux langues. Les résultats indiquent un effet significatif des facteurs *langue* et *genre* sur l'ensemble de ces paramètres. La F0 moyenne est globalement plus élevée en français ; la modulation de F0 est quant à elle globalement plus forte en anglais. Indépendamment de la langue, les locutrices présentent une F0 moyenne plus élevée. En outre, les locutrices ont plus fortement augmenté leur F0 moyenne lors de l'emploi du français et les locuteurs ont réduit plus fortement leur modulation dans cette langue. Ces données suggèrent l'existence de normes vocales relatives au genre différentes dans les deux langues étudiées.

## ABSTRACT

**A study of fundamental frequency in female and male English/French bilingual speakers.**
The present study deals with the productions of English/French bilingual speakers in a reading task and in semi-spontaneous speech. Mean fundamental frequency, F0 range and F0 standard-deviation were measured in both languages. Results show a significant effect of *gender* and *language* on all these parameters. Overall, mean F0 was higher in French while F0 modulation was stronger in English. Regardless of language, female speakers exhibited higher F0 than males. Moreover, the increase of mean F0 in French was larger in female speakers. On the other hand, the decrease of F0 modulation in French was stronger for male speakers. These data support the idea of language- and gender-specific vocal codes, to which bilingual speakers seem to adapt.

MOTS-CLES : fréquence fondamentale, intonation, bilinguisme, voix et genre, variations inter-langues.

KEYWORDS: fundamental frequency, intonation, bilingualism, voice and gender, cross-language variation.

# 1 Introduction

Les différences vocales entre femmes et hommes ont fait l'objet de multiples études phonétiques au cours des dernières décennies. Ces études se focalisent en grande partie sur la fréquence fondamentale (F0), considérée avec les fréquences de résonance comme un paramètre décisif dans ce


10.21437/JEP.2018-21

qui constitue une voix en voix de femme ou en voix d'homme. S'il existe diverses études qui comparent les F0 moyennes de locuteur·ice·s de différentes langues (p. ex. Traunmüller, Eriksson 1995), peu d'études ont pris comme objet les variations intra-individuelles de bilingues lors du passage d'une langue à l'autre et analysé les éventuelles différences genrées dans ces variations. Nous suggérons que l'étude de ces variations permet non seulement de s'éloigner d'une vision statique de la F0, dans laquelle celle-ci est présentée comme une caractéristique essentielle des locuteur·ice·s ou comme principalement due à leur anatomie, pour en adopter une vision dynamique qui permet d'intégrer des questions de différences culturelles de normes de genre et de performances du genre.

Au niveau acoustique, la F0 des voix d'hommes se situe généralement dans des fréquences plus basses que celles des voix de femmes (Boë et al., 1975). Ces différences acoustiques sont en partie dues aux différences sexuées qui émergent lors de la puberté dans les appareils phonatoires. Les taux de testostérone, de progestérone et d'estrogènes varient en fonction de la sexuation du corps et favorisent un développement de plis vocaux plus massifs dans les corps de sexe masculin que dans ceux de sexe féminin (Kahane, 1978 ; Abitbol et al., 1999), ce qui explique en partie pourquoi les plis vocaux des hommes vibrent généralement à une fréquence plus basse que ceux des femmes. À côté des hormones, d'autres facteurs tels que l'âge (Honjo, Isshiki, 1980) ou la consommation de cigarettes (Matar, 2016) peuvent modifier la masse des plis vocaux et ainsi provoquer un abaissement ou une élévation de la F0.

Dans la voix, les facteurs anatomiques et sociaux sont cependant inextricables. Il est désormais établi que la voix participe à la construction sociale des identités de genre (Arnold, 2015 ; Pépiot, 2014b). Chaque locuteur·ice a un appareil phonatoire d'une forme donnée (qui joue sur la F0 et les fréquences de résonance produites), mais fait un usage de cet appareil en fonction de son genre. Une voix n'est ainsi jamais uniquement le reflet d'une anatomie, mais aussi le résultat d'une performance genrée : les femmes mobilisent certaines pratiques articulatoires afin de produire des voix relativement aigues et claires, et les hommes en utilisent d'autres afin de produire des voix relativement graves et sombres (Arnold, 2016).

Dans la littérature, la plage de variation et les modulations de F0 sont elles aussi souvent décrites comme indexant des catégories de genre : les femmes utiliseraient des plages de variations plus étendues que les hommes et moduleraient plus (p. ex. Austin, 1965 ; Lakoff, 1975, p. 56). Ceci a cependant été contredit par certaines études portant sur l'anglais étatsunien. Par exemple, Henton (1989 ; 1995) a montré que si on utilise une échelle de mesure en demi-tons, c'est-à-dire une mesure logarithmique qui reflète la perception humaine des variations de hauteur, alors les différences entre femmes et hommes s'effacent. Pépiot (2014a), en utilisant le même procédé de mesure que Henton a cependant trouvé que les locutrices françaises modulaient plus que les locuteurs français. Ces différences de résultats suggèrent qu'il existe des pratiques genrées différentes concernant la modulation chez les locutrices et locuteurs étatsunien et français.

Qu'en est-il alors des locuteur·ice·s multilingues ? Comment s'adaptent-elles/ils aux normes genrées des différentes langues parlées ? En quoi ces normes diffèrent-elles pour les femmes et les hommes ? Ces questions n'ont pour l'instant fait l'objet que de peu d'attention.

Différentes études réalisées sur des locuteur·ice·s bilingues ont montré qu'en fonction de la langue parlée, ces dernier·e·s vont varier leur F0 moyenne (Altenberg, Ferrand, 2006 ; Lee, Van Lanker Sidtis, 2017) ainsi que leur plage de variation de F0 (Mennen et al., 2012). Par exemple, l'étude d'Altenberg et Ferrand (2006) montre que des locutrices bilingues russes L1 / anglais L2 tendent à parler avec une F0 plus basse en anglais. Cette analyse a cependant été conduite uniquement sur des productions de locutrices – il est conséquemment impossible de savoir si les variations observées relèvent d'une adaptation aux normes genrées des différentes langues, ou tout simplement de pratiques liées aux langues elles-mêmes, sans considération de genre.

Nous avons donc souhaité, par la présente étude, investiguer les pratiques de locutrices et de locuteurs bilingues anglais L1 / français L2, en mesurant leur F0 dans différentes conditions (lecture de phrases et parole spontanée). Notre hypothèse de départ est la suivante : *les locutrices et les locuteurs bilingues adaptent leurs pratiques vocales aux normes genrées de la langue employée*.

# 2 Méthode

## 2.1 Corpus

La présente étude se base sur l'analyse d'un corpus en anglais et en français collecté lors de deux tâches distinctes. Ces deux tâches ont permis de collecter des séquences de parole lue et de parole semi-spontanée.

La première tâche consistait en une lecture de 12 phrases en anglais (telles que « *When the weather is cold and rainy, I'd rather stay at home.* » ; « *My sister told me she'd come by tomorrow.* » ou encore « *If you do that again, I'll call the police!* ») et 12 phrases similaires en français (« *Quand il fait froid et qu'il pleut, je préfère rester chez moi.* » ; « *Ma soeur m'a dit qu'elle allait passer demain.* » ; « *Si tu refais ça, j'appelle la police !* » ; etc.).

La deuxième tâche, qui a permis de collecter de la parole semi-spontanée, consistait en une narration d'un évènement passé – les dernières vacances. La narration en anglais a été amorcée par l'énoncé « *Tell me about your last vacation* », et celle en français par l'énoncé « *Parlez-moi de vos dernières vacances* ».

## 2.2 Participant·e·s

Six locutrices et six locuteurs bilingues anglais L1 / français L2 ont pris part à cette étude. Les participant·e·s sont originaires du nord-est des États-Unis et vivent en région parisienne depuis plusieurs années. Tou·te·s font état d'une pratique quotidienne de la langue française et d'un niveau d'aisance dans cette langue supérieur ou égal à 3, sur une échelle allant de 0 à 5, via un questionnaire inspiré par celui de Grosjean (2013). Nous reprenons également à notre compte la définition du bilinguisme proposé par cet auteur.

Les participant·e·s étaient âgé·e·s de 29 à 54 ans (*SD*=7,6 ans) au moment des enregistrements, avec une moyenne d'âge de 41,8 ans pour les femmes et de 40 ans pour les hommes. Aucun·e n'était fumeur·euse et ne présentait de troubles de la parole. Une clé USB a été offerte en échange de la participation à la présente étude.

## 2.3 Procédure d'enregistrement

Les enregistrements se sont déroulés dans une chambre anéchoïque avec un enregistreur numérique *Edirol R09-HR* de marque *Roland*. Chaque session d'enregistrement comprenait les tâches détaillées dans la section 2.1 : en premier lieu, la lecture des phrases avec un débit de parole moyen (deux lectures pour chaque item), puis la narration portant sur les dernières vacances. Afin de neutraliser de possibles biais liés à l'ordre d'emploi des deux langues (voir Altenberg, Ferrand, 2006), la moitié des locuteur·ice·s a commencé par effectuer ces tâches en langue française, et l'autre moitié en langue anglaise.

## 2.4 Analyse des données

L'analyse acoustique du corpus a été effectuée à l'aide du logiciel *Praat* (Boersma, 2017). Les paramètres suivants ont été relevés pour chacune des phrases ainsi que pour la parole spontanée :

- F0 moyenne.
- Plage de variation de F0, qui correspond à l'écart entre la fréquence la plus basse et la fréquence la plus haute atteinte à l'intérieur d'une unité linguistique donnée.
- Écart-type de F0, qui constitue le paramètre le plus en mesure de rendre compte de la modulation de F0, en particulier lors de l'étude de longues séquences de parole continue.

Ces données ont été obtenues en créant pour chaque phrase/discours un fichier *Pitch* sur Praat, puis en collectant les valeurs dans la fenêtre *Pitch info*. La plage de variation de F0 ainsi que l'écart-type ont été mesurés en Hertz mais également en demi-tons. Cette échelle est en effet particulièrement pertinente car elle rend compte de la variation de hauteur perçue (Henton, 1995).

Les données ainsi recueillies ont ensuite fait l'objet de tests statistiques de type ANOVA, dans le but de tester l'influence des facteurs « langue parlée » et « genre des locuteur·ice·s ».

# 3 Résultats

## 3.1 Phrases lues

La F0 moyenne des locutrices et des locuteurs sur les phrases lues, en anglais et en français, est présentée dans le tableau 1 ci-après.

| | **F0 moyenne - Phrases lues (Hz)** | | |
|---|---|---|---|
| **Loc.** | **Anglais** | **Français** | **% diff. FR/AN** |
| **F1** | 195 | 211 | **+8,28** |
| **F2** | 224 | 234 | **+4,29** |
| **F3** | 176 | 192 | **+8,68** |
| **F4** | 201 | 218 | **+8,37** |
| **F5** | 186 | 205 | **+10,20** |
| **F6** | 187 | 206 | **+10,01** |
| **Moyenne F** | **195** | **211** | **+8,17** |
| **H1** | 113 | 112 | **-1,29** |
| **H2** | 81 | 83 | **+2,63** |
| **H3** | 120 | 121 | **+1,11** |
| **H4** | 106 | 103 | **-3,49** |
| **H5** | 129 | 129 | **-0,10** |
| **H6** | 108 | 119 | **+9,77** |
| **Moyenne H** | **110** | **111** | **+1,33** |

TABLEAU 1 : F0 moyenne en Hertz (Hz) des locutrices et des locuteurs sur les phrases lues (12 x 2 occurrences), en fonction de la langue parlée (anglais ou français).

On constate que les locutrices présentent toutes une F0 moyenne plus élevée en français qu'en anglais. Cette élévation du F0 est de 8,17 % en moyenne, toutes locutrices confondues. Chez les

locuteurs, en revanche, il est difficile de dégager une tendance claire : trois d'entre eux présentent une F0 plus élevée en français, mais l'inverse est vrai pour les trois autres.

Une ANOVA à deux facteurs (« langue parlée » et « genre ») confirme l'influence significative de la langue, avec $F(1,572)=25,566$ et $p<0,0001$, et du genre des locuteur·ice·s, avec $F(1,572)=2897,3$ et $p<0,0001$, sur la F0 moyenne. De plus, on observe une interaction significative entre les deux facteurs ($F(1,572)=17,712$ ; $p<0,0001$). Cela indique que les locutrices et les locuteurs n'ont pas adapté de la même manière leur F0 moyenne en passant d'une langue à l'autre, suggérant ainsi l'existence d'une variation inter-genre sur l'utilisation de ce paramètre acoustique en fonction de la langue parlée.

La plage de variation de F0, en Hertz et demi-tons, ainsi que son écart-type moyen (SD) en Hertz (Hz) et demi-tons (dt) sur les phrases lues sont visibles ci-dessous, dans le tableau 2.

| | **Phrases lues - AN** | | | | **Phrases lues - FR** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Loc.** | **Pl. var. (Hz)** | **Pl. var (dt)** | **SD (Hz)** | **SD** | **Pl. var. (Hz)** | **Pl. var. (dt)** | **SD (Hz)** | **SD (dt)** | **% diff. FR/AN SD (dt)** |
| **F1** | 218,28 | 20,37 | 50,56 | 4,85 | 219,36 | 19,06 | 40,53 | 3,60 | **-25,76** |
| **F2** | 233,25 | 20,21 | 46,37 | 3,79 | 203,33 | 15,92 | 39,46 | 2,97 | **-21,76** |
| **F3** | 166,34 | 16,71 | 32,56 | 3,22 | 165,56 | 14,95 | 25,24 | 2,29 | **-28,90** |
| **F4** | 201,91 | 17,93 | 38,39 | 3,23 | 224,59 | 19,56 | 41,46 | 3,54 | **+9,60** |
| **F5** | 182,75 | 16,61 | 38,00 | 3,45 | 162,96 | 14,00 | 33,85 | 2,88 | **-16,54** |
| **F6** | 173,64 | 16,88 | 28,12 | 2,68 | 213,91 | 18,91 | 33,74 | 2,92 | **+8,88** |
| **Moy. F** | **196,03** | **18,12** | **39,00** | **3,54** | **198,28** | **17,07** | **35,71** | **3,03** | **-14,26** |
| **H1** | 101,92 | 15,30 | 23,61 | 3,51 | 88,56 | 13,96 | 19,00 | 3,00 | **-14,59** |
| **H2** | 54,12 | 10,68 | 9,78 | 2,01 | 53,50 | 10,47 | 10,53 | 2,14 | **+6,27** |
| **H3** | 91,55 | 13,82 | 22,03 | 3,24 | 82,75 | 11,49 | 19,19 | 2,62 | **-19,05** |
| **H4** | 79,84 | 12,50 | 20,13 | 3,16 | 72,43 | 11,86 | 16,91 | 2,79 | **-11,55** |
| **H5** | 94,08 | 11,95 | 24,41 | 3,13 | 78,71 | 10,32 | 18,48 | 2,41 | **-23,25** |
| **H6** | 76,66 | 12,98 | 16,85 | 2,73 | 69,23 | 9,94 | 14,87 | 2,06 | **-24,84** |
| **Moy. H** | **83,03** | **12,87** | **19,47** | **2,96** | **74,20** | **11,34** | **16,50** | **2,50** | **-15,61** |

TABLEAU 2 : Valeurs moyennes de la plage de variation de F0 en Hertz (Hz) et demi-tons (dt) et de l'écart-type (SD) de F0 en Hertz et demi-tons obtenus par les locutrices (F) et les locuteurs (H) sur les phrases lues (12 x 2 occurrences), en fonction de la langue parlée (anglais ou français).

Les données recueillies montrent que plage de variation est réduite lors des séquences en français. Cette réduction est de 11,89 % (en dt) chez les locuteurs et de 14,36 % (en dt) chez les locutrices.

On note également une tendance globale à une plus faible modulation lors de l'emploi du français. Ce phénomène est plus prononcé chez les locuteurs, avec une diminution de 15,61 % de l'écart-type (en dt), contre 14,26 % chez les locutrices.

Une ANOVA à deux facteurs (« langue parlée » et « genre ») a été conduite sur la plage de variation de F0 (en dt). L'analyse confirme un rôle significatif de la langue, avec $F(1,572)=18,823$ et $p<0,0001$, et du genre des locuteur.rices, avec $F(1,572)=340,109$ et $p<0,0001$. Il en va de même sur l'écart-type, exprimé en demi-tons, pour le facteur langue ($F(1,572)=44,087$ ; $p<0,0001$) et pour le facteur genre ($F(1,572)=57,530$ ; $p<0,0001$).

## 3.2 Parole semi-spontanée

Comme expliqué supra, les locutrices et locuteurs ont outre la tâche de lecture de phrases également eu à produire des séquences de parole semi-spontanée. Les F0 moyennes des participant·e·s lors de ces séquences, qui avaient des durées d'une à deux minutes, sont présentées dans le tableau 3.

| | **F0 moyenne - Discours semi-spontané (Hz)** | | |
|---|---|---|---|
| **Loc.** | **Anglais** | **Français** | **% diff. FR/AN** |
| **F1** | 179 | 189 | **+5,47** |
| **F2** | 190 | 195 | **+2,95** |
| **F3** | 167 | 175 | **+4,73** |
| **F4** | 193 | 197 | **+2,13** |
| **F5** | 173 | 177 | **+2,25** |
| **F6** | 184 | 182 | **-0,98** |
| **Moy. F** | **181** | **186** | **+2,76** |
| **H1** | 104 | 105 | **+0,86** |
| **H2** | 74 | 73 | **-1,62** |
| **H3** | 103 | 105 | **+2,43** |
| **H4** | 99 | 99 | **+0,20** |
| **H5** | 121 | 121 | **+0,50** |
| **H6** | 99 | 100 | **+1,52** |
| **Moy. H** | **100** | **101** | **+0,65** |

TABLEAU 3 : F0 moyenne des locutrices et des locuteurs sur le discours semi-spontané, en fonction de la langue parlée (anglais ou français).

Les résultats obtenus vont dans le sens des tendances observées sur les phrases lues. En effet, cinq des six locutrices ont utilisé une F0 plus élevée lorsqu'elles parlaient français. La sixième présente quant à elle des F0 relativement stables dans les deux langues (-0,98 % en français). Toutes locutrices confondues, on observe une augmentation de la F0 moyenne de 2,76 % en français, comparée à l'anglais. Chez les locuteurs, on constate une relative stabilité de la F0 moyenne avec des variations de +0,65 %.

Une ANOVA à deux facteurs (« langue parlée » et « genre ») confirme l'influence significative de la langue ($F(1,476)=7.059$ ; $p<0,01$) et du genre des locuteur·ice·s ($F(1,476)=6062,193$ ; $p<0,0001$) sur la F0 moyenne. A l'instar des phrases lues, on observe ici aussi une interaction entre les deux facteurs ($F(1,476)=3,816$ ; $p=0,0513$), même si celle-ci n'atteint pas le seuil de significativité.

Le tableau 4, ci-après, présente la plage de variation de F0 (en Hz et dt), ainsi que l'écart-type (également en Hz et dt) lors des séquences en parole semi-spontanée.

| Loc. | Discours semi-spontané - AN | | | | Discours semi-spontané - FR | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Pl. var. (Hz) | Pl. var. (dt) | SD (Hz) | SD (st) | Pl. var. (Hz) | Pl. var. (dt) | SD (Hz) | SD (dt) | **% diff. FR/AN SD (dt)** |
| **F1** | 300,32 | 25,41 | 42,62 | 3,94 | 297,94 | 25,30 | 41,65 | 3,54 | **-10,15** |
| **F2** | 309,61 | 25,82 | 34,26 | 3,24 | 305,69 | 25,64 | 36,92 | 3,34 | **+3,09** |
| **F3** | 211,16 | 20,89 | 22,51 | 2,23 | 253,20 | 23,05 | 23,44 | 2,18 | **-2,24** |
| **F4** | 289,14 | 23,41 | 28,16 | 2,52 | 276,55 | 23,57 | 28,38 | 2,46 | **-2,38** |
| **F5** | 300,88 | 25,42 | 40,94 | 3,64 | 306,62 | 25,61 | 42,90 | 3,52 | **-3,30** |
| **F6** | 244,63 | 22,63 | 30,98 | 2,75 | 286,52 | 24,78 | 29,70 | 2,60 | **-5,45** |
| **Moy. F** | **275,96** | **23,93** | **33,25** | **3,05** | **287,75** | **24,66** | **33,83** | **2,94** | **-3,41** |
| **H1** | 224,44 | 28,21 | 34,33 | 4,21 | 177,13 | 28,42 | 26,78 | 3,56 | **-15,44** |
| **H2** | 67,52 | 16,27 | 7,13 | 1,58 | 53,24 | 12,52 | 6,87 | 1,54 | **-2,53** |
| **H3** | 124,70 | 24,24 | 21,91 | 4,36 | 151,39 | 21,68 | 13,37 | 2,09 | **-52,06** |
| **H4** | 155,71 | 24,44 | 20,11 | 3,18 | 151,03 | 24,08 | 7,53 | 2,95 | **-7,23** |
| **H5** | 179,88 | 21,87 | 25,44 | 3,21 | 176,18 | 21,78 | 20,98 | 2,67 | **-16,82** |
| **H6** | 127,53 | 19,71 | 11,87 | 1,94 | 107,71 | 17,79 | 11,75 | 1,92 | **-1,03** |
| **Moy. H** | **146,63** | **22,46** | **20,13** | **3,08** | **136,11** | **21,05** | **14,55** | **2,46** | **-15,85** |

TABLEAU 4 : Valeurs moyennes de la plage de variation de F0 (en Hz et dt) et de l'écart-type de F0 (en Hz et dt) obtenus par les locutrices (F) et les locuteurs (M) en parole semi-spontanée, en fonction de la langue parlée (anglais ou français).

Si l'on s'intéresse aux différences de l'écart-type de la F0 exprimé en demi-tons entre les deux langues (dernière colonne du tableau 4), on constate que tou·te·s les participant·e·s, à l'exception d'une locutrice, ont moins modulé leur fréquence fondamentale en français qu'en anglais. Cependant, on note que cette tendance est nettement plus marquée chez les locuteurs, avec un écart-type de 15,85 % inférieur en français, alors que chez les locutrices, la diminution n'est que de 3,41 %. Ce résultat rejoint ainsi celui observé sur les phrases lues.

Une ANOVA à deux facteurs (« langue parlée » et « genre ») a été effectuée sur le paramètre de l'écart-type. Elle met en évidence un rôle significatif de la langue, avec $F(1,476)=29,353$ et $p<0,0001$, et du genre des locuteur·ice·s, avec $F(1,476)=11,371$ et $p<0,001$. De plus, on observe une interaction significative entre les deux facteurs ($F(1,476)=14,097$ ; $p<0,001$), ce qui indique que les locutrices et les locuteurs n'ont pas adapté de la même manière la modulation de leur fréquence fondamentale en passant d'une langue à l'autre. Ainsi, la modulation de F0 était globalement similaire chez les locutrices et les locuteurs lors de l'emploi de la langue anglaise – 3,05 demi-tons chez les locutrices et 3,08 demi-tons chez les locuteurs. En revanche, lors des séquences en langue française, les femmes ont plus modulé leur F0 que les hommes – l'écart-type des locutrices est de 2,94 demi-tons et celui des locuteurs de 2,46.

# 4 Conclusion / discussion

Comme il a été indiqué dans la section précédente, nous avons trouvé une interaction significative entre les facteurs *langue* et *genre* sur la F0 moyenne en parole lue et sur l'écart-type de F0 en parole spontanée. On observe également une tendance similaire sur la F0 moyenne en parole spontanée, même si l'interaction entre les deux facteurs étudiés n'atteint pas le seuil de significativité. Ceci indique que dans ces contextes, la langue va globalement jouer sur ces paramètres acoustiques, mais de manière différente en fonction du genre des locuteur·ice·s.

L'analyse de la F0 moyenne a montré que celle-ci est globalement plus élevée en français qu'en anglais, indépendamment du genre des locuteur·ice·s. Cependant, si l'on compare les locutrices et les locuteurs, on constate une différence : lors de l'emploi du français, on trouve une augmentation de la F0 moyenne en parole lue chez toutes les locutrices et en parole spontanée chez 5 sur 6 locutrices, alors qu'on retrouve cette augmentation moins régulièrement chez les locuteurs – ces derniers ont utilisé une F0 moyenne similaire dans les deux langues en parole spontanée et seulement 3 des 6 locuteurs l'ont augmentée en français en parole lue. Le fait que les locutrices et les locuteurs n'aient pas adapté de la même manière leur F0 moyenne en passant d'une langue à l'autre peut être interprété comme un indice d'une différence de genre ethnolinguistique dans l'utilisation de ce paramètre acoustique.

Pour ce qui est des modulations, l'écart-type de la F0 exprimé en demi-tons montre que tou·te·s les participant·e·s, à l'exception d'une seule locutrice, ont moins modulé leur F0 en français qu'en anglais. Cette réduction de modulation est plus prononcée chez les locuteurs, avec une diminution de 15,61 % en parole lue et de 15,85 % en parole spontanée, contre respectivement 14,36 % chez les locutrices en parole lue et 3,41 % en parole spontanée. En outre, nous avons remarqué que lors des séquences produites en anglais en parole spontanée, les locuteurs ont autant modulé leur F0 que les locutrices. Ce résultat va dans le sens de l'étude de Henton (1995) qui, en mesurant les modulations de locuteurs féminins et masculins étasuniens en demi-tons, n'a trouvé aucune différence inter-genres significative. En revanche, lors des séquences en français en parole spontanée, on note que les locuteurs modulaient nettement moins leur F0 que les locutrices. Ces résultats confirment ceux de Pépiot (2014a) et suggèrent qu'en français, les modulations de F0 font partie des pratiques vocales qui contribuent à produire de la différence entre le groupe des femmes et celui des hommes et à diminuer les différences à l'intérieur de ces groupes.

L'analyse de ces différents paramètres, ainsi que l'étude croisée des facteurs *langue* et *genre* sur les productions de bilingues apporte ainsi de nouveaux éléments qui n'avaient pas pu être dégagés des études précédentes, comme celles de Altenberg et Ferrand (2006), Lee et Van Lanker Sidtis (2017) ou Mennen et al. (2012). La présente étude révèle notamment que la production d'une voix de femme ou d'homme mobilise des pratiques vocales différentes en fonction de la langue et confirme ainsi que la F0 n'est pas une caractéristique essentielle des locuteur·ice·s, principalement due à l'anatomie de leur appareil phonatoire, mais qu'elle résulte également d'un apprentissage et d'une socialisation en tant que membre d'une catégorie de genre spécifique. Ceci constitue un argument pour s'éloigner d'une conception anatomiste de la F0 que l'on peut retrouver fréquemment dans la littérature phonétique, et pour une plus grande considération des facteurs sociaux dans l'étude de la voix et de la parole.

Parmi les limites de cette étude, on citera notamment le nombre relativement réduit de participant·e·s. Afin de confirmer nos résultats, la présente étude pourrait être répliquée avec un nombre plus important de locuteur·ice·s. D'autre part, nous avons ici uniquement analysé de la parole produite par des bilingues anglais L1 / français L2. Il serait opportun d'étudier également la parole de bilingues français L1 / anglais L2. En effet, l'augmentation de la F0 moyenne en français pourrait éventuellement être liée, dans une certaine mesure, à un stress éprouvé par les locuteur·ice·s lors de la prise de parole dans leur langue seconde – le stress induisant régulièrement une augmentation de F0 (Scherer 1986). La comparaison avec les productions de bilingues français L1 / anglais L2 permettrait d'établir si la prise de parole en langue seconde agit sur la F0.

# Remerciements

Un grand merci à tou·te·s les locutrices et locuteurs ayant pris part à cette étude, ainsi qu'à Juliette Stockman qui a participé à la préparation et à l'analyse de notre corpus.

# Références


ABITBOL J., ABITBOL P., ABITBOL B. (1999). Sex hormones and the female voice. *Journal of Voice* 13, 424–446.

ALTENBERG E. P., FERRAND C. T. (2006). Fundamental frequency in monolingual English, bilingual English/Russian, and bilingual English/Cantonese young adult women. *Journal of Voice* 20(1), 89-96.

ARNOLD A. (2015). Voix et transidentité : changer de voix pour changer de genre ?. *Langage et société* 151(1), 87−105.

ARNOLD A. (2016). Voix. *Encyclopédie critique du genre*, 713−721. Paris : La Découverte.

AUSTIN W. M. (1965). Some social aspects of paralanguage. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 11(1), 31−39.

BOERSMA P., WEENINK D. (2017). Praat: doing phonetics by computer [Logiciel]. Version 6.0.36, publiée le 11 Novembre 2017 sur le site www.praat.org

BOË L.-J., CONTINI M., RAKOTOFIRINGA H. (1975). Etude statistique de la fréquence laryngienne. *Phonetica* 32(1), 1- 23.

GROSJEAN, F. & LI, P. (2013). *The Psycholinguistics of Bilingualism*. Oxford : Wiley-Blackwell.

HENTON C. (1989). Fact and fiction in the description of female and male pitch. *Language & Communication*, 9(4), 299-311.

HENTON C. (1995). Pitch dynamism in female and male speech. *Language & Communication* 15(1), 43-61.

HONJO I., ISSHIKI N. (1980). Laryngoscopic and Voice Characteristics of Aged Persons. *Archives of Otolaryngology* 106(3), 14-150.

KAHANE J. C. 1978. A morphological study of the human prepubertal and pubertal larynx. *American Journal of Anatomy* 151, 11-19.

LAKOFF R. (1975). *Language and Woman's Place*. New York : Harper & Row.

LEE B., VAN LANCKER SIDTIS D. (2017). The bilingual voice: Vocal characteristics when speaking two languages across speech tasks. *Speech, Language and Hearing* 20(3), 174-185.

MATAR N., PORTES C., LANCIA L., LEGOU T., BAIDER F. (2016). Voice Quality and Gender Stereotypes: A Study of Lebanese Women With Reinke's Edema. *Journal of Speech, Language, and Hearing Research* 59(6), S1608−S1617.

MENNEN I., SCHAEFFLER F., DOCHERTY G. (2012). Cross-language differences in fundamental frequency range: A comparison of English and German. *The Journal of the Acoustical Society of America* 131(3), 2249-2260.

PEPIOT E. (2014a). Male and female speech: a study of mean F0, F0 range, phonation type and speech rate in Parisian French and American English speakers. *Proceedings of the 7th International Conference on Speech Prosody*, 305-309.

PEPIOT, E. (2014b). Voix et genre : un état de la question. In Ibrahim, A.H. (éd.), *La langue, la voix, la parole* (pp. 53-86), Paris : CRL.

SCHERER K. R. (1986). Voice, Stress, and Emotion. *Dynamics of Stress*. Boston : Springer

TRAUNMÜLLER H., ERIKSSON A. (1995). The frequency range of the voice fundamental in the speech of male and female adults. Manuscrit : http://www2.ling.su.se/staff/hartmut/f0_m&f.pdf.

# Evaluation automatique de l'intelligibilité de la parole dans le contexte de cancers de la tête et du cou

Imed Laaridh[1] Corinne Fredouille[1]
Alain Ghio[2] Muriel Lalain[2] Virginie Woisard[3]
(1) LIA, Université d'Avignon
(2) Aix-Marseille Univ, CNRS, LPL, UMR 7309, Aix-en-Provence, France
(3) Service ORL, CHU Larrey, URI Octogone-Lordat, Toulouse, France
corinne.fredouille@univ-avignon.fr

## RÉSUMÉ

Dans le contexte de la parole pathologique, l'évaluation perceptive reste la méthode la plus utilisée par les cliniciens et orthophonistes pour mesurer le niveau d'intelligibilité de leurs patients, et ce malgré le caractère subjectif bien connu de ce type d'évaluation. Le travail présenté ici consiste à reproduire la méthode automatique de prédiction de l'intelligibilité, proposée dans une précédente étude, sur un corpus de patients atteints de cancers de la tête et du cou et présentant des troubles de la parole plus ou moins sévères. Ce travail doit permettre (1) de comparer le comportement du système automatique de prédiction du degré d'intelligibilité sur une population de patients différente, (2) de montrer la pertinence, en terme d'application d'outils automatiques, d'un nouveau protocole d'enregistrements des patients, basé sur des logatomes et dédié à l'évaluation de l'intelligibilité. Les résultats expérimentaux obtenus confirment la validité de l'approche automatique (corrélation r=0.84) mais également du protocole utilisé.

## ABSTRACT

**Automatic evaluation of speech intelligibility in the context of head and neck cancers.**

In disordered speech context, and despite its well-known subjectivity, perceptual evaluation has been, and still, the most commonly used method in clinical practice to evaluate the intelligibility level of speech productions of patients. The work presented in this paper consists in reproducing the automatic method of intelligibility prediction, proposed in a previous study, on a corpus of head and neck cancer patients, and presenting more or less severe disordered speech. This work should permit (1) to compare the behavior of the automatic system for predicting the degree of intelligibility in a different patient population, (2) to demonstrate the relevance, in terms of application of automatic tools, of a new protocol of patient recordings, based on logatomes and dedicated to the evaluation of intelligibility. The experimental results obtained confirm the validity of the automatic approach (correlation r = 0.84) but also the one of the protocol used.

MOTS-CLÉS : Traitement automatique de la parole, i-vecteurs, évaluation de l'intelligibilité, troubles de la parole, cancers de la tête et du cou.

KEYWORDS: automatic speech processing, speech disorders, i-vectors, intelligibility assessment, head and neck cancers.


10.21437/JEP.2018-22

# 1 INTRODUCTION

Les troubles de la parole peuvent affecter, en fonction de leurs origines, différentes composantes de la production de la parole : respiration, phonation, résonance et/ou articulations. Différentes mesures ont été étudiées dans la littérature pour évaluer la qualité de la parole telle que l'intelligibilité, la compréhensibilité et la sévérité de la parole pathologique. Par conséquent, de nombreux protocoles d'évaluation, dédiés à la pratique clinique mais également à la recherche, ont été conçus pour regrouper tout ou partie de mesures. Ces protocoles peuvent être dédiés à des troubles spécifiques de la voix - dysphonie - et/ou de la parole - dysarthrie pour les origines neurologiques ou d'autres troubles de la parole comme dans le cas de cancers de la tête et du cou ou de malformations. Ils visent à aider les cliniciens dans leur connaissance des troubles de la parole et de leur évaluation clinique, essentielle pour suivre la progression de la maladie des patients dans le cas d'un traitement ou l'évolution des altérations dans le cas d'une rééducation de la parole. Dans ce contexte, l'évaluation perceptive reste la méthode la plus utilisée dans la pratique clinique malgré ses limites bien documentées telles que la non reproductibilité et la subjectivité.

La perte d'intelligibilité est l'une des plaintes les plus fréquentes rencontrées chez les patients souffrant de troubles de la parole. Pour faire face aux limitations rapportées ci-dessus, les approches automatiques ont été considérées, très tôt, comme des solutions potentielles en vue d'apporter des outils objectifs pour l'évaluation de l'intelligibilité. Dans la littérature, on peut distinguer deux types principaux d'approches : celles directement basées sur des systèmes automatiques de transcription de parole fournissant un taux d'erreurs de transcription de mots comme score d'intelligibilité (Christensen *et al.*, 2012), et celles utilisant des technologies automatiques capables d'extraire de l'information pertinente, injectée, dans un second temps, dans un système de prédiction automatique du degré d'intelligibilité (Middag *et al.*, 2009; Khan *et al.*, 2014). Parallèlement, d'autres approches automatiques axées sur une analyse plus pointue de la parole dysarthrique, dédiées, par exemple, à la détection d'anomalies, ont également montré des résultats probants dans l'évaluation du degré d'intelligibilité (Laaridh *et al.*, 2015).

Le paradigme des i-vecteurs est une approche état de l'art, appliquée avec succès dans les applications de reconnaissance du locuteur (Dehak *et al.*, 2011). Il est prouvé qu'il représente et capture de manière très pertinente les caractéristiques des locuteurs ciblés(Verma & Das, 2015). Ce paradigme a été utilisé et adapté à plusieurs autres contextes tels que la reconnaissance de la langue et même l'évaluation de la qualité de la parole. Dans (An *et al.*, 2015), cette représentation, combinée à un large ensemble de caractéristiques acoustiques, syllabiques et phonotactiques, a été utilisée pour la prédiction automatique des scores UPDRS (Unified Parkinson's Disease Rate Scale : batterie de tests dédiés à l'évaluation des troubles moteurs de la maladie de Parkinson) de production de parole de patients atteints de la maladie de Parkinson, dans le contexte spécifique du défi ComParE Interspeech 2015. Dans (Wang *et al.*, 2016), le paradigme des i-vecteurs a été utilisé comme une normalisation du locuteur et impliqué dans une approche de classification plus complexe, combinant des caractéristiques acoustiques et articulatoires pour la détection automatique de la Sclérose Latérale Amyotrophique (SLA). Dans (Martínez *et al.*, 2015), les i-vecteurs ont été utilisés pour la représentation de segments de mots produits par 15 locuteurs dysarthriques, permettant d'établir des corrélations importantes entre les mesures d'intelligibilité prédites automatiquement et les mesures d'intelligibilité de référence. Enfin, dans (Garcia *et al.*, 2017), les auteurs ont proposé une approche basée sur une distance cosinus entre la représentation i-vecteur d'une production de parole (test) et deux i-vecteurs de référence représentant individuellement de la parole normale et dysarthrique.

Dans un travail précédent (Laaridh *et al.*, 2017), nous avons proposé une approche basée sur le paradigme des i-vecteurs pour la prédiction automatique de plusieurs métriques d'évaluation de la parole dysarthrique comme l'intelligibilité, la sévérité des troubles de la parole, et plus spécifiquement, le degré d'altération de l'articulation. L'approche proposée a été appliquée sur un corpus de 129 locuteurs dysarthriques et témoins et des mesures de corrélation élevées (entre 0,8 et 0,9) ont été atteintes entre les différentes mesures perceptives d'évaluation de la parole et les prédictions automatiques.

Dans ce travail, soutenu par le projet de recherche C2SI financé par l'Institut National du CAncer (INCA), la même approche automatique basée sur des i-vecteurs est utilisée pour la prédiction automatique de la mesure de l'intelligibilité de la parole, dans le contexte particulier de patients atteints de cancers de la tête et du cou (HNC - Head and Neck Cancer). Comparé au travail précédent, l'objectif de ce travail est double. D'une part, il doit permettre d'observer le comportement de l'approche de prédiction automatique dès lors qu'elle est appliquée sur une population de patients différente. En effet, contrairement à la dysarthrie pour laquelle les troubles de la parole peuvent être diffus, divers et, par conséquent, difficiles à localiser, nous avons ici la connaissance, en fonction du cancer du patient, de la localisation de la déficience (langue, palais, larynx, ...) et une information un peu plus "précise" du trouble de parole attendu. D'autre part, les scores de prédiction automatique sont comparés ici à des mesures d'intelligibilité issues d'un protocole original d'enregistrements et d'évaluation perceptive de productions de parole pathologique basé sur la production de pseudo-mots par des patients et des témoins. Il s'agit par conséquent d'étudier et de valider l'utilisation de ce protocole dans le cadre de l'application d'outils automatiques en vue d'une transposition dans le milieu clinique.

# 2 DESCRIPTION DU CORPUS

## 2.1 Population

L'étude actuelle est basée sur le corpus français de parole HNC, enregistré dans le cadre du projet C2SI. Ce corpus comprend des patients atteints de cancers de la tête et du cou (cavité buccale ou oropharyngés) et des locuteurs témoins. Tous les patients ont subi un traitement dédié consistant en une chirurgie, et/ou une radiothérapie, et/ou une chimiothérapie.
Durant le protocole d'enregistrements conçu spécifiquement dans le cadre du projet C2SI, tous les locuteurs ont été invités à enregistrer différentes tâches de production de la parole (voyelles tenues /a/, lecture d'un texte pour l'obtention de parole lue, lecture d'une liste de phrases pour l'étude de la prosodie, description d'images pour l'obtention de parole spontanée, production de pseudo-mots isolés, etc). Le lecteur pourra se référer à (Astesano *et al.*, 2018) pour une description détaillée des enregistrements de ce corpus, des différentes tâches de production de parole réalisées par les locuteurs et de leurs objectifs de recherche.
Dans ce travail, nous concentrons notre attention uniquement sur la tâche de production de pseudo-mots isolés, appelée DAP (pour Décodage Acoustico-Phonétique) dans le reste de l'article. Durant cette tâche, tous les locuteurs devaient prononcer 52 pseudo-mots (incluant 2 essais d'entraînement). Chaque pseudo-mot suivait une structure phonotactique comme suit : $C(C)_1V_1C(C)_2V_2$ où $C(C)_i$ est une consonne isolée ou un groupe consonantique. Les consonnes ont été sélectionnées à partir d'une liste contenant 18 items, les groupes consonantiques à partir de listes contenant 16 ou 32 items et les voyelles à partir de listes contenant 7 ou 8 items en fonction de leur position dans le

pseudo-mot. En utilisant cette méthode combinatoire, environ 95000 pseudo-mots ont été générés. Cet ensemble a ensuite été réduit à environ 90000 éléments après la suppression des pseudo-mots faisant référence à une entrée lexicale de la langue française. Des listes de 52 pseudo-mots prévues pour la tâche de lecture ont ensuite été construites aléatoirement à partir de l'ensemble des 90000 éléments restants tout en respectant un processus de construction commun. Toutes les listes extraites devaient respecter les mêmes règles concernant la distribution des consonnes, des groupes consonantiques et des voyelles. Au total, 85 patients et 41 témoins ont été enregistrés pour cette tâche. Il convient de noter que certains patients n'ont pas terminé la tâche de lecture en raison d'une fatigabilité extrême et ont produit moins de 52 pseudo-mots requis.

## 2.2 Evaluation perceptive de l'intelligibilité

Tous les pseudo-mots prononcés par les locuteurs du corpus HNC ont été transcrits par 40 auditeurs, suivant le protocole décrit dans (Ghio *et al.*, 2018). Vu l'ampleur de la tâche, chaque pseudo-mot a été évalué par 3 d'entre eux. Le choix des auditeurs naïfs présentant un bon niveau d'orthographe a été délibéré afin d'éviter tous effets d'habituation à la parole pathologique (altérations et phénomènes de compensation) bien connus chez les cliniciens et qui peuvent faciliter leur compréhension. Ces auditeurs ont été confrontés à une tâche qui s'apparente à un décodage acoustico-phonétique (d'où le nom de la tâche de production de la parole - DAP) suivi d'une transcription écrite. Au total, 18360 transcriptions orthographiques des pseudo-mots ont été recueillies. L'annotation a été réalisée en utilisant la plate-forme Lancelot-Perceval (Ghio *et al.*, 2003). Chaque auditeur pouvait écouter chaque pseudo-mot jusqu'à 3 fois avant de fournir sa transcription. Pour chaque pseudo-mot, la distance moyenne de Levenshtein a été utilisée pour comparer la suite de phonèmes attendue et les réponses transcrites, compte-tenu des caractéristiques acoustiques distinctives entre phonèmes. La distance moyenne a ensuite été calculée, pour chaque locuteur, sur l'ensemble des pseudo-mots produits oralement, fournissant ainsi, pour chacun d'eux, une mesure d'(in)intelligibilité (les valeurs élevées correspondent à la plus grande distance entre le pseudo-mot attendu et la réponse transcrite et caractérisent donc les locuteurs les moins intelligibles).

# 3 METHODOLOGIE

L'approche automatique utilisée ici, et décrite précédemment dans (Laaridh *et al.*, 2017), repose sur deux étapes. La première étape consiste à projeter chaque énoncé de parole dans le sous-espace de variabilité totale de faible dimension et de représenter ainsi chaque enregistrement associé à un locuteur témoin ou à un patient par un i-vector (Dehak *et al.*, 2011).
La deuxième étape est une régression du sous-espace des i-vecteurs vers l'espace d'intelligibilité (à 1 dimension) nécessaire à notre tâche d'évaluation de cette dernière. La régression par Machines à Vecteurs de Support (RVS) sera utilisée compte-tenu du nombre limité de données annotées disponibles pour l'étude. En effet, malgré le grand nombre de patients et de sujets témoins disponibles, la quantité de données disponible reste limitée par rapport à d'autres applications de traitement automatique de la parole «standard».

## 3.1 Espace de variabilité totale

Le paradigme de l'espace de variabilité totale a d'abord été introduit dans le contexte de la reconnaissance automatique du locuteur. Dans cette approche, un extracteur d'i-vecteurs convertit une séquence de vecteurs acoustiques en un seul vecteur de faible dimension représentant l'ensemble de l'énoncé de parole. Le super-vecteur $s$ dépendant du locuteur et de la session issu de la concaténation des vecteurs de moyennes d'un modèle de mélange gaussien (GMM) est supposé obéir à un modèle linéaire de la forme : $s = m + Tw$
où $m$ est le super-vecteur moyen du modèle de parole générique ou modèle du monde (UBM), $T$ est la matrice de projection de faible rang apprise sur un large ensemble de données par estimation MAP (elle représente le sous-espace de «variabilité totale») et $w$ est une variable latente, appelée "i-vector", ayant une distribution normale $\mathcal{N}(0, I)$. Les algorithmes pour l'estimation de $T$ et l'extraction des i-vecteurs sont décrits dans (Matrouf *et al.*, 2007).

## 3.2 Extraction des i-vecteurs

Une étape de paramétrisation des signaux de parole, nécessaire à tout processus automatique, repose ici sur l'extraction de 19 coefficients cepstraux (LFCC), leurs 19 dérivées premières ($\Delta$) et leurs 11 dérivées secondes ($\Delta\Delta$). Une normalisation de la moyenne et de la variance (MVN) est ensuite appliquée aux paramètres LFCC, valeurs estimées sur les portions de parole de chaque enregistrement détectées à l'aide d'un alignement phonétique automatique contraint par le texte. Un UBM de 512 composantes dépendant du genre et une matrice de variabilité totale $T$ de rang 400 estimée à partir des corpus de parole français Ester 1 & 2, REPERE et ETAPE (7690 sessions de 2906 locuteurs) (Ajili *et al.*, 2016) sont utilisées pour extraire un i-vecteur par enregistrement de parole. Le package LIA_SpkDet de la boîte à outils open source ALIZE (Larcher *et al.*, 2013) est utilisé pour les différents traitements en lien avec les i-vecteurs décrits ci-dessus.

## 3.3 Régression par Machines à Vecteurs de Support (SVR)

Dans $\epsilon$-SVR, l'idée de base est de trouver une fonction qui a, au plus, $\epsilon$ d'écart par rapport aux valeurs de référence cibles pour toutes les données d'entraînement. Quand une telle tâche n'est pas réalisable, des variables de compromis et de relâchement sont introduites pour faire face au problème d'optimisation (Smola & Schölkopf, 2004).
Pour chaque vecteur de test, et compte-tenu des vecteurs d'apprentissage $x_i \in R^{400}$, $i = 1, ..., n$, la fonction de décision est alors :

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b \tag{1}$$

où $\alpha_i$ et $\alpha_i^*$ sont des multiplicateurs de Lagrange, $K$ est la fonction du noyau et $b$ est le biais. Dans ce travail, les noyaux RBF ont été utilisés.
En outre, et compte-tenu des excellentes performances obtenues par la représentation à base d'i-vecteurs dans le domaine de la reconnaissance du locuteur, un processus de validation croisée sur 10 sous-ensembles a été mis en place pour éviter les biais liés à la présence des mêmes locuteurs dans les phases d'apprentissage et de test.

# 4 RESULTATS ET DISCUSSIONS

Pour évaluer les performances de l'approche automatique sur le corpus HNC, les mesures de corrélation de Pearson ($r$) et d'erreur quadratique moyenne (RMSE) ont été calculées entre les scores d'intelligibilité prédits automatiquement et les mesures d'intelligibilité issues de l'évaluation perceptive dans le cadre du DAP (section 2.2).

## 4.1 Prédiction de l'intelligibilité au niveau locuteur

La première expérience que nous avons réalisée consistait en la prédiction automatique de l'intelligibilité de chaque locuteur en utilisant l'ensemble des enregistrements de parole, dans lequel il prononçait les 52 pseudo-mots. Cela signifie que nous disposions d'une quantité importante de données pour estimer le i-vecteur représentant la production acoustique de chaque locuteur incluant les altérations de parole. La figure 1 fournit la mesure d'intelligibilité prédite automatiquement par rapport à l'évaluation perceptive de référence issue du protocole DAP par locuteur. Nous observons que l'approche automatique est capable d'effectuer une bonne séparation entre les patients et les groupes de locuteurs témoins. La pente de régression confirme la capacité du système à détecter et représenter la perte d'intelligibilité mesurée perceptivement par les auditeurs. En effet, les mesures $r$ et RMSE atteignent respectivement 0.84 et 2.339. Ce taux de corrélation est tout à fait cohérent avec les résultats précédents observés sur la tâche de lecture de texte produite par des patients dysarthriques (Laaridh *et al.*, 2017). De plus, la mesure RMSE obtenue est assez faible considérant que l'intervalle de la mesure de référence est caractérisé par une large variation (mesures de référence appartenant à l'intervalle [0,22] pour ce corpus) et sa sensibilité extrême (l'évaluation perceptive est mesurée sur un intervalle discret ; la moindre différence de caractéristiques sur n'importe quel phonème entre la référence et la transcription des pseudo-mots aboutit, au minimum, à une distance de 1 point).

## 4.2 Prédiction de l'intelligibilité au niveau de sous-listes de mots

La pertinence de l'approche automatique, en terme de prédiction de l'intelligibilité sur les pseudo-mots étant confirmée dans la section précédente, nous proposons ici d'étudier la quantité de parole requise pour effectuer une prédiction fiable. En effet, comme mentionné dans la section 2, la fatigabilité des patients peut les contraindre à abandonner un tâche de production de parole dans le cadre d'un protocole d'évaluation. S'assurer que l'ensemble du protocole d'enregistrement est nécessaire au bon fonctionnement des outils automatiques est un aspect majeur. Vérifier que ce dernier pourrait être allégé en est un autre, tout aussi crucial, notamment dans une perspective de pratique clinique. Considérant que chaque locuteur a produit au moins une liste de 52 pseudo-mots, nous avons divisé la liste de ces derniers en 5 sous-listes d'environ 10 mots chacune. Chaque sous-liste représentait environ 7 secondes seulement de parole, ce qui est évidemment très court dans le cadre d'un traitement automatique suivant la tâche visée. En dehors de leur taille, la distribution des pseudo-mots parmi les sous-listes reste totalement aléatoire et ne tient pas compte de leur structure ou de leurs phonèmes de composition. Chaque sous-liste a été assignée à une mesure d'intelligibilité perceptive en faisant la moyenne des distances mesurées sur les pseudo-mots (section 2.2) la composant. Au total, 623 sous-listes ont été extraites représentant les 126 locuteurs. De manière identique à l'expérimentation précédente, une validation croisée de 10 sous-ensembles a été mise en œuvre au niveau du locuteur pour éviter les biais résultant de l'utilisation de sous-listes produites par le même locuteur dans les

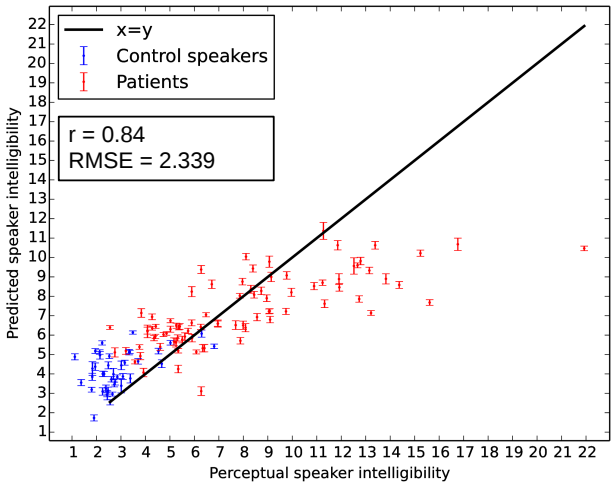



FIGURE 1 – Mesures d'intelligibilité par locuteur, issues de la prédiction automatique sur l'ensemble des 52 pseudo-mots et des mesures perceptives (DAP). Droite de pente 1 (noir) donnée pour indication.

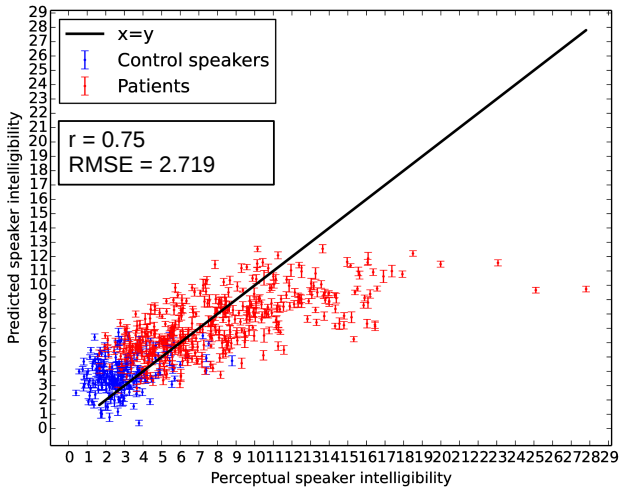



FIGURE 2 – Mesures d'intelligibilité par sous-liste (incluant ∼10 pseudo-mots), issues de la prédiction automatique et des mesures perceptives (DAP). Droite de pente 1 (noir) donnée pour indication.

phases d'apprentissage et de test.

La figure 2 fournit la mesure d'intelligibilité prédite automatiquement par rapport à l'évaluation perceptive de référence issue du protocole DAP pour chacune des sous-listes produites par les locuteurs. Nous observons que le caractère discriminant de l'approche automatique entre patients et locuteurs de contrôle reste conservé et que les 10 sous-listes de pseudo-mots utilisées étaient de taille suffisante pour l'approche automatique de détection de la perte d'intelligibilité pour les patients.

## 4.3 Discussions

Une hypothèse clé faite dans l'expérience précédente était que les différentes sous-listes, extraites aléatoirement, étaient équivalentes et portaient la même information du point de vue de l'intelligibilité. Cependant, l'observation des mesures de corrélation entre les prédictions automatiques et

TABLE 1 – Mesures de corrélation de Pearson ($r$) et d'erreur quadratique moyenne (RMSE) entre scores d'intelligibilité prédits et perceptifs en fonction des meilleures et des pires sous-listes.

| | $r$ | RMSE |
|---|---|---|
| Sélection des meilleures sous-listes | 0.87 | 1.685 |
| Sélection des pires sous-listes | 0.58 | 4.278 |

les évaluations perceptives calculées sur les 5 sous-listes extraites par locuteur montre des valeurs variant de 0,72 à 0,80. Cette variabilité signifie que les sous-listes, et donc les pseudo-mots, ont été considérées différemment par le système de prédiction automatique et l'approche à base d'i-vecteurs. Partant de cette observation, le choix des sous-listes à utiliser pour la tâche de prédiction semble avoir un impact non négligeable.

Afin de mettre en évidence ce comportement, la table 1 présente les meilleures (pire) mesures $r$ et RMSE qui pourraient être obtenues si nous considérions seulement les sous-listes qui minimisent (maximisent) les erreurs de prédiction. Nous observons que la mesure de corrélation pourrait atteindre jusqu'à 0,87 et la RMSE seulement 1,685 si les sous-listes utilisées pour l'évaluation sont bien choisies. En revanche, la mesure $r$ descend à 0,58 si les sous-listes utilisées contiennent des mots «moins significatifs». Même si, dans ce cas, la perte de performance est importante, cette valeur peut être considérée comme un seuil en termes de mauvaises prédictions pour l'approche automatique. Comme reporté plus haut, le potentiel discriminant d'un pseudo-mot en termes d'intelligibilité peut varier ostensiblement. Par conséquent, en plus de la méthode sophistiquée de conception de la liste initiale de 90000 pseudo-mots (voir section 2.1), nous pouvons facilement supposer qu'une méthodologie plus réfléchie pour composer les sous-listes de pseudo-mots pourrait entraîner des gains encore plus importants en terme de précision de la prédiction automatique.

Cette observation peut être très utile lors de la mise en œuvre de nouveaux protocoles pour l'évaluation automatique / perceptive des troubles de la parole. En effet, la plupart des tests d'évaluation sont substantiels et nécessitent beaucoup d'efforts de la part du patient (et par conséquent de la fatigue potentielle) et de l'auditeur évaluateur. Une sélection plus pertinente des unités linguistiques pourrait s'avérer extrêmement utile pour réduire de manière drastique les efforts à fournir.

# 5 CONCLUSIONS ET PERSPECTIVES

Cet article étudie une approche automatique pour la prédiction de l'intelligibilité de la parole basée sur le paradigme des i-vecteurs et des modèles de régression à base de machines à support de vecteurs. Cette approche a été appliquée sur un protocole de lecture dédié de pseudo-mots produits par des locuteurs souffrant de cancers de la tête ou du cou. Une corrélation élevée (r=0,84) a été obtenue entre les mesures d'intelligibilité prédites automatiquement sur chaque locuteur et celles issues de l'évaluation perceptive basée sur le DAP dès lors que les 52 pseudos-mots sont considérés dans un seul enregistrement de parole. Par ailleurs, l'approche s'est avérée stable et robuste au manque de données puisque $r$ = 0,75 a été atteint en utilisant seulement 20% environ de la production de parole de chaque locuteur (pseudo-mots $\sim$10). Les résultats de cette dernière expérimentation a permis de montrer l'effet des sous-listes utilisées pour l'évaluation, et par conséquent de la pertinence des pseudo-mots qui les composent en termes de prédiction de l'intelligibilité. Les travaux futurs étudieront les informations portées par chaque pseudo-mot et l'impact de son contenu phonétique sur l'évaluation de l'intelligibilité, du point de vue perceptif et automatique.

# Références


AJILI M., BONASTRE J.-F., BEN KHEDER W., ROSSATO S. & KAHN J. (2016). Phonetic content impact on forensic voice comparison. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*, p. 210–217 : IEEE.

AN G., BRIZAN D. G., MA M., MORALES M., SYED A. R. & ROSENBERG A. (2015). Automatic recognition of unified parkinson's disease rating from speech with acoustic, i-vector and phonotactic features. In *Proceedings of Interspeech'15*, Dresden, Allemagne.

ASTESANO C., BALAGUER M., FARINAS J., FREDOUILLE C., GAILLARD P., GHIO A., GIUSTI L., LAARIDH I., LALAIN M., LEPAGE B., MAUCLAIR J., NOCAUDIE O., PINQUIER J., PONT O., POUCHOULIN G., PUECH M., ROBERT D., SICARD E. & WOISARD V. (2018). Carcinologic speech severity index project : A database of speech disorders productions to assess quality of life related to speech after cancer. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2018)*, Myazaki, Japan.

CHRISTENSEN H., CUNNINGHAM S., FOX C., GREEN P. & HAIN T. (2012). A comparative study of adaptive, automatic recognition of disordered speech. In *Proceedings of Interspeech'12*, Portland, USA.

DEHAK N., KENNY P. J., DEHAK R., DUMOUCHEL P. & OUELLET P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, **19**(4), 788–798.

GARCIA N., OROZCO-ARROYAVE J. R., D'HARO L., DEHAK N. & NÖTH E. (2017). Evaluation of the neurological state of people with parkinson's disease using i-vectors. In *Proceedings of Interspeech'17*.

GHIO A., ANDRÉ C., TESTON B. & CAVÉ C. (2003). Perceval : une station automatisée de tests de perception et d'évaluation auditive et visuelle. *Travaux interdisciplinaires du Laboratoire parole et langage d'Aix-en-Provence (TIPA)*, **22**, 115–133.

GHIO A., LALAIN M., GIUSTI L., POUCHOULIN G., ROBERT D., FREDOUILLE C., LAARIDH I. & WOISARD V. (2018). Une mesure d'intelligibilité par décodage acoustico-phonétique de pseudo-mots dans le cas de parole atypique. In *Journées d'Etude sur la Parole (JEP), Aix-en-Provence, France, Juin 2018*.

KHAN T., WESTIN J. & DOUGHERTY M. (2014). Classification of speech intelligibility in parkinson's disease. *Biocybernetics and Biomedical Engineering*, **34(1)**, 35–45.

LAARIDH I., BEN KHEDER W., FREDOUILLE C. & MEUNIER C. (2017). Automatic prediction of speech evaluation metrics for dysarthric speech. In *Proceedings of Interspeech'17*, p. 1834–1838.

LAARIDH I., FREDOUILLE C. & MEUNIER C. (2015). Automatic detection of phone-based anomalies in dysarthric speech. *ACM Transactions on accessible computing*, **6**(3), 9 :1–9 :24.

LARCHER A., BONASTRE J.-F., FAUVE B. G., LEE K.-A., LÉVY C., LI H., MASON J. S. & PARFAIT J.-Y. (2013). Alize 3.0-open source toolkit for state-of-the-art speaker recognition. In *Proceedings of Interspeech'13*, p. 2768–2772.

MARTÍNEZ D., LLEIDA E., GREEN P., CHRISTENSEN H., ORTEGA A. & MIGUEL A. (2015). Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace. *ACM Transactions on Accessible Computing (TACCESS)*, **6**(3), 10.

MATROUF D., SCHEFFER N., FAUVE B. G. & BONASTRE J.-F. (2007). A straightforward and efficient implementation of the factor analysis model for speaker verification. p. 1242–1245.

MIDDAG C., MARTENS J.-P., VAN NUFFELEN G. & DE BODT M. (2009). Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on Advances in Signal Processing*, **2009**(1), 1–9.

SMOLA A. J. & SCHÖLKOPF B. (2004). A tutorial on support vector regression. *Statistics and computing*, **14**(3), 199–222.

VERMA P. & DAS P. K. (2015). i-vectors in speech processing applications : a survey. *International Journal of Speech Technology*, **18**(4), 529–546.

WANG J., KOTHALKAR P. V., CAO B. & HEITZMAN D. (2016). Towards automatic detection of amyotrophic lateral sclerosis from speech acoustic and articulatory samples. In *Proceedings of Interspeech'16*.

# Evaluation de la compréhensibilité et conservation des fonctions prosodiques en perception de la parole de patients post traitement de cancers de la cavité buccale et du pharynx

Olivier Nocaudie[1] Corine Astésano[1,] Alain Ghio[2], Muriel Lalain[2], Virginie Woisard[1, 3]

(1) Octogone-Lordat (E.A. 4156), Université de Toulouse, France
(2) Aix-Marseille Univ, CNRS, LPL, UMR 7309, Aix-en-Provence, France
(3) C.H.U. de Toulouse-Rangueil, France
nocaudie@univ-tlse2.fr

## RESUME

Dans le cadre du projet "Carcinologic Speech Severity Index", cette étude propose l'évaluation perceptive de la parole de 45 locuteurs (37 patients ayant subi des traitements pour des cancers de la cavité buccale et/ou du pharynx, 8 témoins). 66 auditeurs naïfs ont effectué 4 tâches perceptives investiguant, d'une part, la conservation de fonctions prosodiques (groupements syntaxiques, focus pragmatique, modalisation d'énoncés) chez ces patients et, d'autre part, leur compréhensibilité (paradigme de Sentence Verification). Les résultats indiquent une conservation globale des fonctions prosodiques testées chez les patients. En parallèle, ces résultats montrent que le paradigme de Sentence Verification permet de discriminer les locuteurs en fonction de leur compréhensibilité perçue et que cette tâche présente une corrélation forte avec l'évaluation subjective de la sévérité de la parole par des praticiens hospitaliers. Ces résultats sont discutés en regard du rôle joué par la prosodie dans l'adaptabilité communicative des locuteurs en cas de déficience articulatoire.

## ABSTRACT

**Comprehensibility evaluation and conservation of prosodic function in the speech of patients after Head and Neck carcinologic treatment**

In the framework of the Carcinologic Speech Severity Index, the present study proposes to perceptually evaluate speech of 45 speakers (37 patients, with treatment for Head and Neck cancer, 8 controls). 66 naïve listeners fulfilled 4 tasks, investigating on the one hand the conservation of 3 prosodic functions in these patients' speech (syntactic grouping, pragmatic focus, expression of modality) and their intelligibility/comprehensibility using a Sentence Verification Task on the other hand. Results indicate a global conservation of the tested prosodic functions in these patients' speech. The Sentence Verification paradigm results however help discriminate groups of speakers based on their perceived comprehensibility and show higher correlation with subjective evaluation of speech severity by medical practitioners. The results are discussed with regards to the communication adaptative role of prosody in the case of articulatory impairments.

MOTS-CLES : Parole carcinologique, perception, compréhensibilité, prosodie

KEYWORDS: Carcinologic speech, perception, intelligibility, prosody



10.21437/JEP.2018-23

# 1 Introduction

Les traitements des cancers des voies aérodigestives supérieures (VADS), du fait de leur caractère souvent mutilant malgré les reconstructions anatomiques, peuvent avoir un impact important sur les fonctions de communication. En effet, les organes atteints par ces pathologies sont impliqués dans la production de la parole, notamment au niveau articulatoire (cavité buccale, langue, pharynx). En conséquence, les traitements de ces cancers (i.e. chirurgie, et/ou chimiothérapie) perturbent la production de la parole des patients et, corollairement, sa réception par leurs interlocuteurs.

En raison de la diminution de la mortalité liée aux traitements des cancers, l'évaluation des résultats fonctionnels du traitement carcinologique est une préoccupation majeure des équipes soignantes pour envisager un pronostic fonctionnel concernant la qualité de vie à venir du patient, pour ajuster les protocoles thérapeutiques et améliorer la vie communicationnelle des patients.

## 1.1 Carcinologic Speech Severity Index (C2SI)

Le projet Carcinologic Speech Severity Index (ci-après C2SI) vise à établir un indice automatique mesurant l'impact des traitements carcinologiques des VADS (*i.e*, ici, des cancers localisés dans la cavité buccale et/ou le pharynx) sur la qualité de la communication des patients. A cette fin, les acteurs du projet C2SI étudient un même corpus original, composé de nombreuses tâches linguistiques (tenues de voyelles, lecture de texte, production de pseudo-mots, description d'images, etc.) recueilli auprès de 135 sujets (Astésano *et al.*, 2018), que ce soit sur le plan acoustique, (Sicard *et al.*, 2017), en perception de la parole (Ghio *et al*., 2017 ; Nocaudie *et al.*, 2017) ou bien en traitement automatique (Lahrid *et al*., 2017).

Pour mettre au point le C2SI, un index automatique évaluant la dégradation de la communication parlée post cancer, il nous est nécessaire d'envisager la parole de ces patients du point de vue de sa réception par les auditeurs humains. D'une part, le recours à l'évaluation perceptive de la parole permet d'obtenir des points de comparaison (Hermes, 1998) pour estimer la validité perceptive d'un indice automatique, *e.g.* C2SI, même si l'holisme de ce procédé présente certaines limites. D'autre part, le handicap lié aux séquelles des traitements des cancers des VADS concerne particulièrement la réception de cette parole altérée. En effet, les difficultés d'articulation dues à la modification des organes de la parole conduiraient à une perte d'intelligibilité du message, *i.e.* de l'aisance des auditeurs à récupérer le détail phonétique du message et parallèlement à une altération de la compréhensibilité du message, *i.e.* la capacité de l'auditeur à interpréter le message indépendamment de ses altérations (Woisard et al..2013). Dans ces termes, l'intelligibilité renvoie au traitement des informations de bas niveaux, tandis que la compréhensibilité évoque des stratégies de haut niveau : par exemple, reconstruction du sens, compensations de l'auditeur ou jugements de plausibilité à partir des connaissances préalables.

## 1.2 Prosodie et compréhensibilité

En parole, les indices prosodiques se trouvent à l'interface des autres niveaux linguistiques et remplissent de nombreuses fonctions participant à l'encodage comme au décodage des messages. La prosodie permet notamment de marquer la structuration des énoncés en indiquant les frontières des unités, remplissant alors une fonction syntaxique, de moduler la modalité d'un même énoncé par des variations d'indices prosodiques (forme du contour, intensité, etc.) ou encore de faire ressortir l'information centrale d'un message par son marquage accentuel (par exemple,

production d'accent initial emphatique). La prosodie joue alors un rôle pragmatique et informationnel (Di Cristo, 2012).

Au-delà de ces fonctions communicatives, la production cohérente des indices prosodiques d'une langue participe à la fluence des énoncés et à la gestion des aspects temporels de la parole. Par ses fonctions, et ses interactions avec les autres niveaux linguistiques, la prosodie est donc essentielle à la compréhensibilité des messages.

Certains modèles voient d'ailleurs la prosodie comme faisant partie de stratégies compensatoires/palliatives à des altérations segmentales, *i.e.* les locuteurs tendraient à amplifier les indices prosodiques dans leurs productions pour compenser une perte d'intelligibilité/compréhensibilité, due au bruit ambiant ou à des difficultés de prononciation (théorie de l'adaptabilité de la parole à des fins d'optimisation de la communication ; Lindblom, 1990).

### 1.3 Cadre de l'étude

Dans le cadre du projet C2SI, cette étude propose ainsi d'évaluer d'une part, la compréhensibilité des locuteurs au moyen d'une tache de compréhension de phrases (*Sentence Verification Task* ou SVT, d'après Pisoni & Dedina, 1986) et d'autre part le degré de conservation de trois fonctions prosodiques à l'interface avec la syntaxe, l'expression des modalités et la pragmatique. En effet, malgré une atteinte articulatoire consécutive au(x) traitement(s) dont ils ont bénéficié, les patients parviennent-ils à produire de manière satisfaisante les indices prosodiques, *i.e.* les fonctions prosodiques sont-elles conservées chez ces locuteurs, malgré une perte d'intelligibilité/compréhensibilité souvent constatée par les cliniciens ?

### 1.4 Hypothèses

Théoriquement, les traitements de ces cancers des VADS ne sont pas supposés impacter la production du flux laryngé et des indices prosodiques, même si certains types de traitement peuvent conduire à une rigidification des tissus au-delà de la zone traitée et, conséquemment à une altération fonctionnelle du larynx. Ainsi, nous posons l'hypothèse que les locuteurs du C2SI obtiendront des résultats satisfaisants aux évaluations perceptives pour ce qui concerne la conservation des fonctions prosodiques malgré ces risques périphériques : il sera difficile de distinguer les groupes patients/témoins sur la seule base de leurs scores, notamment durant les tâches où l'information segmentale est négligeable (syntaxe & modalité). Cependant, nous pensons que l'atteinte articulatoire des patients aura un impact sur le temps de réponse des auditeurs en perception. En effet, l'altération de la parole des patients se traduirait par des difficultés de traitement par les auditeurs. Il est donc en revanche hautement probable que le groupe patient obtienne des résultats moins bons que ceux du groupe contrôle au test de compréhensibilité (SVT), en raison de ces difficultés de traitement de la parole altérée par les auditeurs.

## 2 Aspects méthodologiques

### 2.1 Recueil du matériau linguistique

Une cohorte de 135 locuteurs enregistrés dans le service d'oncoréhabilitation de l'Oncopole à Toulouse (94 patients / 41 sujets contrôles) a participé à un recueil de données de parole constitué de l'ensemble des tâches de C2SI (Astésano *et al*. 2018). Pour cette étude, nous avons sélectionné

une sous-partie du corpus C2SI comprenant les productions de 45 locuteurs (37 sujets patients et 8 sujets contrôles). Ces productions correspondent à la première phase d'enregistrement de C2SI.

Notre étude porte donc sur quatre tâches de C2SI, trois évaluant de grandes fonctions prosodiques (groupements syntaxiques, marquage du focus pragmatique, marquage de la modalité) et une destinée à l'évaluation de la compréhensibilité (Sentence Verification Task).

### *2.1.1 Prosodie : Groupement syntaxique*

Les locuteurs devaient résoudre une ambiguïté syntaxique par des moyens prosodiques dans des syntagmes composés de deux noms et un adjectif (Astésano, Bard & Turk (2007). L'ambiguïté réside dans la portée de l'adjectif qui peut alternativement qualifier le second nom du syntagme, ou bien les deux noms :

a. [Les tomates][et les **oignons rouges**]

b. [Les **tomates** et les **oignons**][**rouges**]

Les énoncés étaient présentés en modalité écrite et le groupement syntaxique attendu marqué par des indications visuelles. Chaque locuteur a enregistré le même ensemble de 26 énoncés (13 scripts différents en ordre aléatoire, par bloc de groupement syntaxique * 2 conditions syntaxiques).

### *2.1.2 Prosodie : Marquage du Focus pragmatique*

Cette tâche demandait aux locuteurs de marquer le focus pragmatique en mettant en valeur l'information importante d'un énoncé par des moyens prosodiques (Astésano *et al*., 2004; Magne *et al*, 2005. Corpus adapté pour les patients par Aura, 2012). Suite à l'écoute d'une question indiquant quel focus produire « *Où as-tu vu un canard, dans le jardin ou dans la cour* ? » *vs.* « *Qu'as-tu vu dans le jardin, un cochon ou un canard* ? », les sujets devaient lire la réponse en marquant le focus contrastif de manière adéquate : (« *j'ai vu un canard* ***dans le jardin*** » *vs.* (« *j'ai vu un* ***canard*** *dans le jardin* »). Chaque locuteur a enregistré le même ensemble de 20 énoncés (10 contenus lexicaux * 2 focus pragmatiques, présentés selon un ordre aléatoire).

### *2.1.3 Prosodie : Expression de la Modalité*

Lors de cette tâche les locuteurs devaient produire des énoncés identiques au niveau lexical, *e.g.* « *tu prends la voiture* », en les modulant au niveau prosodique, afin de transmettre trois modalités différentes (assertion, interrogation, ordre). Les énoncés étaient présentés en modalité écrite, et la modulation prosodique demandée était marquée par un symbole de ponctuation ('.' ; ' ?' ; ' !'). Chaque locuteur a enregistré un même ensemble de 30 énoncés (10 scripts en ordre aléatoire *3 blocs de modalité).

### *2.1.4 Compréhensibilité : Sentence Verification Task*

Une liste de 150 couples de phrases vraies/fausses et dont le caractère vrai ou faux ne peut être vérifié qu'à partir de la dernière unité lexicale a été constituée (*e.g.* « *La poule pond des œufs* » vs. « *La poule pond des fruits* »). Certains énoncés sont adaptés des propositions de Pisoni & Dedina (1986), d'autres de celles de Zumbiehl (2010), les derniers couples de phrases ont été constitués par les membres du projet C2SI. Chaque locuteur a enregistré un ensemble de 50 énoncés tirés des listes SVT.

Notre corpus pour cette étude représente donc un ensemble de 5670 énoncés [45 sujets * (26 items syntaxe + 30 items modalité + 20 items focus + 50 items SVT)]

## 2.2 Tâches en perception

Suite à ce recueil, une cohorte de 66 auditeurs (18-31 ans, 54 femmes/12 hommes) naïfs (*i.e.* non habitués à la parole altérée pour éviter l'effet des processus de haut niveau dans la perception de la parole, Ohala, 1986) ont procédé à l'évaluation perceptive de ces trois fonctions prosodiques ainsi que de la compréhensibilité des locuteurs. Les auditeurs étaient indemnisés (10 €, une heure de passation) pour leur participation à l'expérience qui s'est déroulée sur la plateforme « Comportement Cognition Usages » de l'université Toulouse–Jean Jaurès.

L'ensemble des tâches a été scripté pour le logiciel PERCEVAL (André et al., 2003), de manière à ce que chaque énoncé produit lors du recueil soit évalué par au moins trois auditeurs différents. Chaque tâche décrite lors du recueil équivalait à un bloc expérimental en perception, au sein duquel les items étaient randomisés. Les blocs étaient eux-mêmes présentés dans un ordre aléatoire. Enfin, notons qu'avant chaque bloc, une phase d'entraînement était accomplie par les auditeurs pour se familiariser avec les tâches.

### *2.2.1 Tâche de perception des frontières syntaxiques (SYN)*

Lors de la tâche SYN, il était demandé aux participants d'indiquer la structure syntaxique comprise après écoute d'un stimulus (ex : [Les tomates][et les **oignons rouges**] vs [Les **tomates** et les **oignons**][**rouges**]). Une image rappelant les cas possibles était présentée à l'écran lors de l'écoute. Chaque participant évaluait 50 à 55 énoncés (en fonction du groupe).

### *2.2.2 Tâche de perception du focus pragmatique (FOC)*

Pour construire cette tâche de perception du focus pragmatique, nous avons associé des questions impliquant la production du focus pragmatique (les mêmes enregistrements que ceux utilisés lors du recueil du corpus) avec des réponses enregistrées par les patients. Nous avons ainsi créé des dialogues question/réponse en manipulant la congruité/incongruité du dialogue pour ce qui concerne la production du focus pragmatique : « *Où as-tu vu un canard, dans le jardin ou dans la cour* ? », puis « *j'ai vu un CANARD dans le jardin* » (réponse inappropriée) *vs.* « *j'ai vu un canard dans le JARDIN* » (réponse appropriée). Après écoute de chaque dialogue, les participants devaient indiquer si la réponse à la question était appropriée ou non. Chaque participant évaluait un ensemble de *ca*.40 dialogues.

### *2.2.3 Tâche de perception de la modalité (MOD)*

Durant le bloc MOD, les participants écoutaient un énoncé avant de devoir en indiquer la modalité comprise parmi trois possibles (interrogation, assertion, ordre). Chaque participant évaluait *ca.* 61 énoncés.

### *2.2.4 Sentence Verification Task (SVT)*

Pour la SVT, les énoncés vrai/faux recueillis lors des enregistrements étaient présentés auditivement aux participants qui devaient indiquer si la phrase entendue était plutôt vraie (« *la poule pond des œufs* ») ou plutôt fausse (« *la poule pond des fruits* »). Chaque participant devait vérifier la véracité de *ca.* 100 phrases.

Pour chacune de ces tâches, nous avons calculé un score par locuteur qui équivaut à la moyenne de chacune des évaluations perceptives reçues durant le test. Dans la mesure où chaque item a été évalué trois fois à travers les auditeurs, ce score est compris entre 0 et 3. Par ailleurs, les temps de réaction des auditeurs ont été calculés (du début de l'écoute de l'énoncé à la réponse, temps auquel est soustrait la durée de l'énoncé) et moyennés pour chacun des locuteurs, afin qu'ils soient adjoints à nos résultats. Ainsi, chaque locuteur du C2SI a obtenu 4 scores différents (un par tâche) associés à des temps de réaction moyens des auditeurs.

# 3 Résultats

## 3.1 Différences intergroupes : Témoins *vs.* Patients

Dans la TABLE 1, nous indiquons pour chacun de nos groupes (Témoins vs. Patients) des indicateurs de statistiques descriptives afin d'observer des tendances de performances d'une part et de variations intragroupes d'autre part.

| **Tâche** | **Variable** | **Témoins** | | | **Patients** | | |
|---|---|---|---|---|---|---|---|
| | | Moyenne | Médiane | Ecart type | Moyenne | Médiane | Ecart type |
| **SVT** | Scores | 2,91 | 2,92 | 0,04 | 2,36 | 2,75 | 0,66 |
| | TR | 1049 | 1046 | 119 | 2750 | 1699 | 2078 |
| **FOC** | Scores | 2,59 | 2,66 | 0,21 | 2,33 | 2,40 | 0,35 |
| | TR | 987 | 965 | 413 | 1891 | 1820 | 796 |
| **SYN** | Scores | 2,21 | 2,22 | 0,39 | 1,81 | 1,69 | 0,39 |
| | TR | 2418 | 2363 | 277 | 2428 | 2219 | 563 |
| **MOD** | Scores | 2,17 | 2,37 | 0,53 | 1,77 | 1,79 | 0,44 |
| | TR | 1754 | 1654 | 278 | 2262 | 2256 | 424 |

TABLE 1 : Indices de variabilité individuelle par groupes. Les scores sont compris entre 0 et 3. Les temps de réaction (TR) sont donnés en millisecondes.

Ainsi, le groupe « Témoins » présente des scores moyens plus élevés ainsi que des temps de réaction moyens plus court que le groupe « Patient ». Cependant, la différence la plus flagrante entre les deux groupes est illustrée par les écarts types du groupe « Patient », beaucoup plus élevés, qui témoignent d'une grande dispersion des données au sein du groupe (tant dans les scores moyens reçus que dans les temps moyens de réaction relevés), notamment pour ce qui concerne la tâche de SVT. Ceci étant dit, les scores moyens obtenus par les deux groupes aux tâches de modalité et de syntaxe sont décevants, plus particulièrement pour le groupe Témoin.
Afin de confirmer ces différences intergroupes en termes de scores moyens et temps de réaction moyens, nous avons conduit des Anovas à un facteur (Témoins vs Patient). Les résultats de ces tests indiquent des différences statistiquement significatives en ce qui concerne les **scores reçus** à la tâche de focus ($F(1, 773) = 12.72$, $p.<.001$), de jugement de la modalité ($F(1, 1137) = 22.50$, $p.<001$), de groupement syntaxique ($F(1, 1038) = 25.43$, $p.<.001$) ainsi qu'à la SVT ($F(1, 1825) = 71.25$, $p.<.001$). A propos des **temps de réaction**, nos tests ont indiqué des différences statistiquement significatives pour toutes les tâches [focus ($F(1, 773) = 52.84$, $p.<.001$) ; modalité ($F(1, 1137) = 37.57$, $p.<001$) ; SVT ($F(1, 1825) = 65.81$, $p.<.001$)], excepté pour la tâche de groupement syntaxique ($F(1, 1038) = 0.007$, *ns*).

### 3.2 Corrélation entre résultats des tests perceptifs et évaluations cliniques

Finalement, nous avons voulu savoir si les évaluations accordées par notre jury d'auditeurs naïfs reflétaient les résultats obtenus par les locuteurs par rapport aux évaluations cliniques. En effet, chaque locuteur a été évalué par 5 professionnels de santé sur deux indices : un indice de sévérité et un indice d'intelligibilité. Dans la mesure où les résultats de nos corrélations étaient très proches pour les deux indices, nous ne présentons dans la TABLE 2 que les corrélations avec l'indice de sévérité.

| Tâche | Facteur | r | p value |
|---|---|---|---|
| SVT | Score moyen | 0.81 | <.001 |
| | Temps de réaction | -0.76 | <.001 |
| FOC | Score moyen | 0.56 | <.001 |
| | Temps de réaction | -0.67 | <.001 |
| MOD | Score moyen | 0.44 | <.05 |
| | Temps de réaction | -0.66 | <.001 |
| SYN | Score moyen | 0.53 | <.001 |
| | Temps de réaction | -0.29 | ns |

TABLE 2 : résultats des tests de corrélation de Pearson entre l'indice de sévérité et le score moyen obtenu par les locuteurs/le temps de réaction des auditeurs. Les valeurs de l'indice de sévérité sont comprises entre 0 (altération très sévère) et 10 (aucune altération).

Ainsi, la corrélation entre score moyen à la SVT et sévérité estimée du locuteur révèle un lien positif fort entre les deux variables ($r = 0.81$, $p.<.001$). Les tests de corrélation avec les scores obtenus aux autres tâches montrent également des liens positifs, mais plus mesurés, s'échelonnant de $r = 0.44$ (MOD) pour le plus faible à $r = 0.56$ (FOC) pour la plus modérément forte.
Par conséquent, la SVT semble constituer une tâche pertinente pour estimer la compréhensibilité du locuteur.

A propos des corrélations entre temps de réaction moyen des auditeurs et indice de sévérité, comme attendu, chaque test indique une relation négative modérément forte à forte [$-0.76 < r < -0.66$] pour l'ensemble des tâches, indiquant alors que plus l'indice de sévérité estimé est bas (*i.e.* plus l'atteinte du locuteur est sévère) plus le temps de réaction de l'auditeur lors de la tâche est élevé, exception faite de la tâche de syntaxe, où le score de corrélation est faible (et non significatif).

# 4 Discussion & Conclusion

Nos résultats présentent plusieurs tendances différentes en fonction des tâches effectuées en perception.
Les résultats de la SVT sont les plus nets, les Patients ayant une moins bonne performance que les Témoins et provoquant des temps de réaction plus longs chez les auditeurs. Cela dit, le groupe Patient montre, particulièrement pour cette tâche, une grande variabilité, les meilleurs atteignant une performance proche de certains Témoins. De plus, les résultats de cette tâche présentaient une corrélation forte avec l'évaluation perceptive de la sévérité des locuteurs par des praticiens hospitaliers. La SVT constituerait ainsi une tâche adéquate pour évaluer la compréhensibilité des locuteurs. Cependant, une étude plus poussée des différences interindividuelles (base de données cliniques C2SI) au sein du groupe patient nous semble nécessaire pour expliquer la variabilité observée (*cf.* TABLE 1).
En prosodie, les résultats étaient moins tranchés. La tâche de focus, a présenté la meilleure corrélation avec l'évaluation clinique, les meilleurs scores moyens et les meilleurs temps de

réaction pour les deux groupes, bien que les patients aient été significativement moins performants que les Témoins. Pour affiner ces résultats, il conviendrait de passer sur le plan acoustique pour observer la réalisation des indices acoustique des accents initiaux emphatiques ainsi que leur localisation dans l'énoncé.

A propos de la tâche de syntaxe et, dans une moindre mesure, de la tâche de modalité, les témoins se sont peu distingués des patients (différences plus réduites que dans les autres tâches, voire non significatives). De plus, les scores moyens obtenus se rapprochent des seuils de hasard. Ceci laisse penser que les locuteurs enregistrés ont pu avoir des difficultés à produire les effets attendus, par exemple, en raison de la difficulté de la tâche de groupement syntaxique qui reste peu naturelle, y compris pour les locuteurs sains (voir Astésano *et al.*, 2007, pour une discussion). En ce qui concerne la production des modalités, la question se pose de savoir si les contours intonatifs produits par les locuteurs étaient suffisamment prototypiques des trois modalités (question, ordre, assertion) pour être reconnus comme tels par les auditeurs. Ces questions constituent des perspectives à développer en prolongement de ce travail, en utilisant des métriques de comparaison des formes prosodiques (Nocaudie, 2016) pour s'assurer que la réalisation phonétique des groupements syntaxiques ou des modalités sont prototypiques.

Finalement, l'autre tendance nette de nos résultats est constituée par la différence de temps moyens de réaction obtenus qui sont systématiquement plus longs pour le groupe patient. Cela pourrait révéler des difficultés de traitement de l'information chez nos auditeurs naïfs, qui n'ont pas l'habitude d'être exposés à de la parole altérée et n'avaient pas la possibilité de réécouter plusieurs fois les stimuli. En SVT, nous pourrions admettre que les locuteurs étaient parfois incapables de répondre car ils n'arrivaient pas à récupérer l'information lexicale. En prosodie, cela semble s'échelonner en fonction des tâches. La tâche FOC ressemble à la SVT (et suit les mêmes tendances), mais dans le domaine prosodique : le locuteur doit maintenir une compréhensibilité suffisante des unités lexicales sur l'énoncé entier et produire les indices prosodiques pertinents. En tâche de MOD ou SYN, les temps de réaction des Témoins et des Patients tendent à se rejoindre : le contour prosodique pourrait suffire seul à accomplir la tâche mais des altérations articulatoires pourraient perturber la perception des contours par les auditeurs (perte de fluence, par exemple). Ainsi, la tâche de focus pourrait représenter une tâche cumulant évaluation de la compréhensibilité et des aspects prosodiques de la parole : évaluation d'indices locaux (réalisation d'IA) et globaux (mesures de fluence, débit, intensité, étendue, etc).

Ce travail doit se poursuivre dans plusieurs directions : en intégrant l'intégralité des locuteurs du corpus C2SI et en exploitant les métadonnées associées (base clinique) dans nos modèles statistiques ainsi que sur le plan acoustique, pour pouvoir décider s'il faut inclure/exclure certaines tâches utilisées dans un outil automatisé type C2SI.

Enfin, cette étude souligne, parfois encore en pointillés, des potentiels en termes de production d'événements prosodiques dans le groupe patient. Si les analyses acoustiques montraient effectivement une bonne conservation des fonctions communicatives et structurels de la prosodie, alors faudrait-il envisager de les exploiter, chez des patients peu intelligibles (indices prosodiques de rythme) et sensibiliser l'entourage des patients à ces enjeux.

# Remerciements

Financement n°2014-135 de l'Institut National pour le CAncer (INCA) de novembre 2014, "Sciences Humaines et Sociales, Épidémiologie et Santé Publique". Porteur : Pr Virginie Woisard.

# Références


ANDRE C., GHIO A., CAVE C., & TESTON B. (2003). PERCEVAL: a Computer-Driven System for Experimentation on Auditory and Visual Perception. In *Proceedings of XVth ICPhS* (p.1421‑1424). Barcelone, Espagne

ASTÉSANO C., BARD, E. G. & TURK, A. (2007). Structural influences on initial accent placement in French. *Language and Speech*, 50(3), 423-446.

ASTESANO C., BALAGUER, M., FARINAS, J., FREDOUILLE, C., …, WOISARD, V. (2018). Carcinologic Speech Severity Index Project: A Database of Speech Disorders Productions to Assess Quality of Life Related to Speech After Cancer. *Proceedings of the 11th Language Resources and Evaluation Conference*, papier accepté, Miyazaki, Japon

AURA K. (2012). *Protocole d'évaluation du langage fondé sur le traitement de fonctions prosodiques : étude exploratoire de deux patients atteints de gliomes de bas grade en contexte péri-opératoire*. Doctoral dissertation, Université Toulouse le Mirail-Toulouse II.

DI CRISTO A. (2012). Le pouvoir de la prosodie ou la revanche de Cendrillon. In Baqué L. & Estrada M. *L'Homme Communiquant*, CIPA, pp.95-114, 2012.

GHIO A., LALAIN M., GIUSTI L., ROBERT D., POUCHOULIN G., …, WOISARD V. (2017). Du décodage acoustico-phonétique pour mesurer l'intelligibilité de locuteurs atteints de troubles de production de la parole, *7emes Journées de Phonétique clinique*, Paris, 29-30 june

HERMES D. J. (1998). Measuring the Perceptual Similarity of Pitch Contours. *Journal of Speech, Language and Hearing Research*, *41*, 73–82.

LAARIDH I., KHADER B. W., FREDOUILLE C. & MEUNIER C. (2017). Automatic Prediction of Speech Evaluation Metrics for Dysarthric Speech, *Interspeech'17*, Stockholm, Sweden.

LINDBLOM B. (1990) Explaining Phonetic Variation: A Sketch of the H&H Theory. In: Hardcastle W.J., Marchal A. (eds) *Speech Production and Speech Modelling*, vol 55. Springer, Dordrecht

MAGNE C.; ASTÉSANO C.; LACHERET-DUJOUR A.; MOREL M.; ALTER K. & BESSON M. (2005) On-line processing of "pop-out" words in spoken french dialogues. *Journal of Cognitive Neuroscience, 17 (5),* 740-756.

PISONI D. & DEDINA M. (1986) "Comprehension of Digitally Encoded Natural Speech using a Sentence Verification Task: a first report" in *Research on Speech Perception. Progress Report N°12*, Indiana University

OHALA J. (1986). Phonological evidence for top-down processing in speech perception. In *Invariance and variability in speech processes* (Erlbaum, p. 386‑397). Hillsdale: Perkell & Klatt

NOCAUDIE O (2016). *Imitation et contrôle prosodique dans l'entraînement à la remédiation phonétique : évaluation, mesure et applications pour l'enseignant de langue étrangère*. Doctoral dissertation, Université de Toulouse.

NOCAUDIE O., ASTESANO C. & WOISARD V. (2017). Conservation des fonctions prosodiques post traitement des cancers de la cavité buccale et du pharynx, *7emes Journées de Phonétique clinique*, Paris-29-30 june

SICARD E., MAUCLAIR J. AND WOISARD V. (2017). Etude de paramètres acoustiques des voix de patients traités pour un cancer ORL dans le cadre du projet C2SI, *7emes Journées de Phonétique clinique*, Paris, 29-30 june

WOISARD V., ESPESSER R., GHIO A. & DUEZ D. (2013). De l'intelligibilité à la compréhensibilité de la parole, quelles mesures en pratique clinique ?. *Rev Laryngol. Otol. Rhinol*., 134 (1), 27-33.

ZUMBIEHL O. (2010). Evaluation perceptive des dysphonies par la Sentence Verification Task. Mémoire d'Orthophonie (dir. : Cavé C. & Ghio A.). Université Aix-Marseille.

# L'effet de la fréquence lexicale sur les réalisations des rhotiques en Écosse

Monika Pukli[1]

(1) LiLPa EA 1339, 22 rue R. Descartes, 67100 Strasbourg, France
mpukli@unistra.fr

## RESUME

Cet article présente les résultats d'une étude auditive et acoustique des réalisations du phonème /r/ en position de coda dans l'anglais parlé en Écosse, avec comme objectif de tester le lien entre la fréquence lexicale et les formes non rhotiques. Bien que la lénition des rhotiques en Écosse partant d'une vibrante roulée et évoluant vers une approximante accompagnée de réalisations de plus en plus vocalisées devrait être sensible à la fréquence lexicale, les données relevées en parole spontanée chez dix-huit locuteurs ne montrent pas de corrélation significative entre celle-ci et les formes non rhotiques. Par conséquent, l'article propose de s'interroger d'une part sur la cohérence méthodologique, et d'autre part sur la possibilité d'interprétation de la fréquence à différentes échelles.

## ABSTRACT

**Lexical frequency effects on rhotic realisations in Scotland**

This paper presents results from an auditory and acoustic study of coda /r/ realisations in Scottish English related to potential frequency effects on non-rhotic occurrences. Although lenition from the Scottish trill towards approximants and more vocalic realisations should be sensitive to word frequency, findings from spontaneous speech in 18 speakers do not show a significant correlation between non-rhotic forms and lexical frequency. Consequently, we raise questions on the coherence across methodological practices on the one hand, and on the interpretation of observed frequency effects on different levels, on the other.

MOTS-CLES : fréquence lexicale, théories basées sur l'usage, anglais écossais

KEYWORDS: lexical frequency, usage-based grammar, Scottish English



10.21437/JEP.2018-24

# 1 Introduction

Cette étude s'interroge sur le rôle de la fréquence lexicale dans le changement phonologique dans le contexte de la dérhoticisation en Écosse de nos jours. La variation phonétique des formes rhotiques en anglais écossais, qui peut être conditionnée par des facteurs liés à l'usager et à l'usage, fait partie d'une lénition graduelle particulière car, mené à terme, le changement suppose une réinterprétation des contraintes phonotactiques. Cette lénition est déterminée par le profil sociolinguistique du groupe d'usagers ainsi que par les caractéristiques phonologiques du mot (y compris, en fonction du cadre théorique, la fréquence lexicale). C'est sur ce dernier point que l'étude acoustique et auditive se concentre ici.

En 1929, dans sa thèse intitulée *Relative frequency as a determinant of phonetic change*, Zipf a établi une corrélation entre certains changements phonologiques et le rang de fréquence du phonème dans la langue. Depuis, mais surtout à l'instar des propositions de Chen et Wang (1975), l'étude du changement s'intéresse tout particulièrement à la diffusion lexicale des formes innovantes et il semble que, selon le type de changement, l'actuation puisse toucher soit en premier lieu les mots fréquents, soit d'abord les mots rares. Au-delà d'une description fidèle de la diffusion des innovations dans la population, les approches sociocognitives ont également élaboré une représentation mentale basée sur la fréquence lexicale des items (phonèmes, morphèmes, collocations, etc.). Plusieurs modèles plus ou moins 'extrêmes' de la Théorie des Exemplaires se basent aujourd'hui sur l'hypothèse selon laquelle la fréquence d'un mot a une influence directe sur sa représentation mentale (voir Ernestus & Baayens, 2011 pour un survol de ces modèles).

La dérhoticisation en Écosse fait référence aux résultats de différentes études qui mettent en relation l'âge des locuteurs et leur tendance à recourir à des réalisations vocaliques, voire absentes du phonème /r/ en position de coda. Même s'il n'existe à notre connaissance pas d'étude de véritable grande ampleur[1], plusieurs enquêtes indépendantes ont démontré l'usage plus ou moins étendu de séquences non-rhotiques dans différentes villes d'Écosse, et les chercheurs s'accordent sur le fait qu'il existe une évolution tendant vers la perte de la rhoticité.

Du point de vue phonologique, ce processus peut représenter une évolution graduelle au niveau phonétique menant à un remaniement du système phonologique à terme (au contraire d'un changement abrupt, pour ceux qui souscrivent à cette dichotomie, où une variante remplace une autre sans réalisations intermédiaires). Le changement graduel peut être perçu dans des formes très hétérogènes, consonantiques et vocaliques, se situant sur un continuum de réalisations intermédiaires entre le [r] traditionnel qui est attesté encore aujourd'hui et l'absence totale du phonème. Mais surtout, la dérhoticisation peut également être conçue comme une lénition ayant pour origine cette vibrante roulée alvéolaire remontant à il y a plus de cent ans, passant par deux réalisations dominantes 'en compétition', vibrante battue vs. approximante, et arrivant à une réalisation entièrement vocalique ou zéro (Stuart Smith et al, 2014). Selon le point de vue adopté, on s'attend

[1] Johntson (1997) examine notamment 91 locuteurs d'Edimbourg, et Jauriberry (2016) présente 161 locuteurs de cinq régions d'Ecosse.

ainsi soit à une absence de l'effet de fréquence lexicale (changement graduel), soit au contraire à un lien fort entre fréquence et formes non rhotiques (lénition).

De notre point de vue, la perte de rhotiques consonantiques s'inscrit dans l'évolution de lénition et nous souhaitons donc tester l'hypothèse d'une corrélation entre forme non rhotique et mot fréquent. Ainsi, dans un schéma simplifié de la Théorie des Exemplaires, la lénition d'un segment dans la production quotidienne et répétée produite par un relâchement persistant de l'effort pour atteindre la cible articulatoire devrait induire un changement dans l'exemplaire, ce qui à son tour se manifesterait dans une production plus rapidement modifiée et donc *plus répandue dans les mots fréquents*.

# 2 Corpus et méthodologie

Les données analysées proviennent du corpus phonologique du programme PAC enregistré au début des années 2000. Dix-huit locuteurs écossais ont été étudiés en parole continue spontanée : huit hommes et dix femmes d'âges variant entre 18 et 82 ans, d'un milieu socio-économique allant d'ouvrier à ingénieur. Ils sont tous nés et ont grandi dans la ville d'Ayr même ou dans ses environs immédiats. La variation en matière de rhoticité a déjà été relevée chez certains locuteurs de ce corpus (Jauriberry et al, 2012) mais pas sur la totalité de l'échantillon.

Afin de tester l'effet potentiel de la fréquence lexicale sur les variantes phonétiques nous avons utilisé le corpus BNC, le corpus oral PAC de la variété locale étant trop petit. Le BNC, avec ses 100 millions de mots lui-même, n'est pas très conséquent par rapport à nos standards actuels, et pour cette raison les lemmes venant du corpus entier ont été pris en compte (Kilgarriff, 1998), mélangeant oral et écrit des années 1980-1990. Les noms propres (items orthographiés avec une majuscule) et les chiffres ont été ignorés, et la valeur forfaitaire 790 a été assignée aux mots avec moins de 800 occurrences dans le BNC.

Nous avons recueilli toutes les occurrences de /r/ en position de coda non suivi de voyelle dans les conversations enregistrées en équilibrant leur nombre entre locuteurs ayant un échange de 20 minutes et ceux qui parlent moins longtemps, c'est-à-dire entre 25 et 40 mots par sujet. Le repérage des occurrences a systématiquement été effectué à partir de la fin et vers le début de l'entretien. De ces données initialement retenues, nous avons enlevé les réalisations non utilisables à l'issue de l'analyse acoustique et les résultats sont ainsi basés sur 566 mesures.

Il est important de noter que les mots grammaticaux ne font pas partie de la sélection étudiée. En effet, ces mots ont une fréquence exceptionnellement haute mais sont articulés de manière très approximative dans les séquences sans accent en anglais. Leurs réductions prépondérantes, y compris naturellement celle du phonème /r/, auraient potentiellement une influence sur les résultats, or les modifications dues au rythme du discours et au tempo de l'élocution constituent des effets qui touchent tous les phonèmes et pas uniquement /r/. Par ailleurs, certaines occurrences de mots lexicaux sont également affectées par un tempo très élevé ; ces réalisations de réduction de syllabe

entière ont été retenues pour l'analyse (et sont donc autant d'occurrences de non-rhoticité dans nos résultats).

Les réalisations des rhotiques ont d'abord été analysées de manière auditive sur deux sessions par le même évaluateur (l'auteure). Une analyse acoustique des formants a ensuite été utilisée pour les occurrences où l'étude auditive s'est avérée non conclusive. Cette analyse a suivi les conventions de la littérature[2] tout en acceptant les formes ambiguës comme étant rhotiques, compensant ainsi d'une part la pauvreté relative d'indices sûrs dans le spectre acoustique pour certaines des formes phonétiques que peuvent prendre les rhotiques en anglais, et d'autre part le nombre incertain d'occurrences ayant une articulation consonantique invisible sur l'image acoustique à cause de l'achèvement du geste lingual survenant après l'arrêt de la phonation (Lawson et al, 2008). Ce procédé représente un risque de surestimation de formes rhotiques qui nous semble être préférable au cas contraire. La non-rhoticité globale observée est ainsi de 42 % des occurrences totales.

# 3 Résultats et discussion

Les variables retenues pour l'appréciation des résultats ont été les suivantes : la valeur de la fréquence lexicale (forme brute et plusieurs variables transformées), l'accentuation (syllabe accentuée vs. non accentuée), l'environnement segmental (position finale avant pause (*anymore*), position finale pré-consonantique (*better than*) et position médiane pré-consonantique (*mermaid, escort*)), ainsi que l'âge, le sexe et le statut socio-économique des locuteurs.

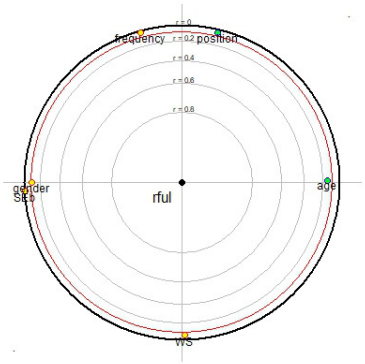



FIGURE 1: L'interaction entre variables, projetée par une analyse en composante principale focalisée (variables explicatives en partant de la gauche dans le sens de l'aiguille d'une montre: sexe, profil socio-économique, fréquence lexicale, position (avant pause ou non), âge, accent syllabique).

[2] Pour les critères de segmentation des 14 formes rhotiques les plus courantes, voir Jauriberry (2016 : 218), pour une étude récente sur l'interaction entre acoustique, gestes articulatoires et perception de la structure formantique des rhotiques écossaises, voir Lawson et al (2018).

Nous avons d'abord tenté d'étudier l'interaction des variables explicatives en vue de réaliser une régression logistique. L'analyse en composante principale focalisée (voir Figure 1 ci-dessus) effectuée par la commande *fpca* de la librairie *psy* avec R montre que les variables présentes dans notre tableau n'ont pas de corrélation positive entre elles. Au centre du cercle se trouve la rhoticité (la variable à expliquer) et nous nous intéressons à des points-variable regroupés dans l'image autour de la fréquence. On peut constater que la fréquence lexicale est bien-entendu indépendante des caractéristiques des usagers tel que l'âge, le profil socio-économique et le sexe. Par ailleurs, toutes les variables à l'exception de l'âge sont en deçà du cercle rouge et seront donc vraisemblablement non significatives (au seuil de 5%).

Si l'on compare la distribution de la variable fréquence (données non transformées) avec celle de l'âge en fonction de la présence et de l'absence de /r/ (Figure 2 ci-dessous), on voit clairement l'effet de l'âge (à droite) alors que l'étendue de la fréquence lexicale des mots ayant des formes non rhotiques est très semblable à celle des mots rhotiques (à gauche).

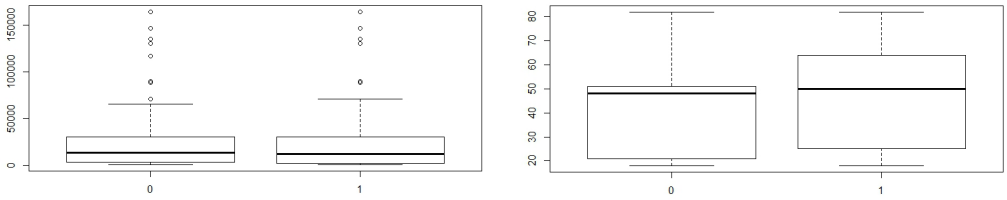



FIGURE 2: La distribution des formes non rhotiques (0) et rhotiques (1) selon la fréquence (à gauche) et selon l'âge (à droite)

Une régression logistique au seul effet de fréquence donne une valeur de p non significative aussi bien pour la fréquence dans sa forme brute (p=0.65922) que dans sa forme logarithmique (p=0.185) par rapport aux occurrences des formes non rhotiques. De même, lorsque les observations sont recodées avec une variable de rang de fréquence allant de 0.2 (le mot le plus rare dans le corpus, *bard*) à 208 (le mot le plus fréquent, *year*), il n'y pas de lien entre la (non) réalisation de /r/ et la fréquence de mot (p=0.20188). Les effets combinés de fréquence + accent lexical, et de fréquence + position de /r/ dans le mot ne sont pas significatifs non plus (respectivement 0.1188 et 0.227).

Le seul facteur phonologique ayant un lien avec la rhoticité est celui de l'accent lexical (X-squared = 5.8302, df = 1, p-value = 0.01575) : les formes non rhotiques apparaissent plus souvent dans les syllabes non accentuées. Il convient cependant d'ajouter qu'il s'agit d'un codage non empirique ; l'accentuation est notée selon la forme isolée du mot sur deux niveaux : absence vs présence d'accent.

Dans un second temps, la variable de la fréquence a été reformulée en variable catégorielle avec quatre valeurs allant de mot rare à très fréquent : 790-2935, 2936-12613, 12614-30594 et 30595-

163930, correspondant aux quatre quartiles de la variable numérique. Cette fois-ci une corrélation légèrement significative émerge entre la catégorie de fréquence et la rhoticité des mots (p=0.04321) si l'on compare les mots rhotiques et non rhotiques.

Comment interpréter ces résultats ? D'abord se pose toujours la question de la méthodologie utilisée (et de l'échantillon sur lequel les mesures et calculs sont basés) : combien de locuteurs, combien de tokens, quel type de données (lecture de mot, lecture de phrase, lecture de texte, échange cadré par une activité précise, échange en présence ou non de l'enquêteur, etc. sont autant de facteurs ayant une influence possible sur les résultats), quelles analyses statistiques. A ces questions d'ordre général s'ajoutent celles en lien avec la fréquence :

- la variable de la fréquence lexicale est-elle établie avec le logarithme de la valeur, en valeur absolue, ou encore avec une recatégorisation en 'rare-moyen-fréquent' ? Quels sont les paramètres de la recatégorisation ?
- la distribution de la fréquence lexicale des mots, connue pour être non normale, est-elle un facteur à prendre en compte dans l'analyse statistique ?
- la fréquence d'un mot donné provient-elle d'un corpus oral de la variété locale (et par conséquent plus petit) ou d'un corpus 'généraliste' de la variété ?
- la liste de fréquence de ce corpus de référence est-elle lemmatisée, et si non, comment les formes sont-elles comptées ?
- au-delà de la fréquence des mots, celle des phonèmes et des cooccurrences de phonèmes peut-elle être pertinente ?

Il est rare que tous ces choix et caractéristiques soient clairement documentés, et peu nombreuses sont les études pouvant être reproduites.

Puis, l'interprétation des résultats revient sur la nature du phénomène étudié. De notre point de vue, la perte de rhotiques consonantiques s'inscrit dans l'évolution de lénition décrite plus haut, or les résultats observés ne sont pas ceux attendus. Il est possible que l'effet de fréquence ne se montre que sur certaines échelles et que cela soit masqué dans l'effet global.[3] La transformation des données brutes de la fréquence lexicale est nécessaire à cause de la distribution de la valeur qui suit une loi de puissance au lieu d'une loi normale. Mais la variable obtenue par transformation logarithmique n'est statistiquement pas liée à la rhoticité.

On ne peut pas passer ici en revue les études tentant à lier fréquence et rhoticité en anglais. A titre d'exemple, Nagy et Irwin (2010) ont trouvé un lien faible entre rhoticité et fréquence en anglais parlé à Boston, cependant, l'étude la plus pertinente sur l'anglais écossais (Jauriberry 2016) n'a pas observé d'effet de fréquence dans la variation des réalisations rhotiques. Cela est autant plus troublant que l'auteur a effectué une recatégorisation de la fréquence.

La représentation en exemplaires qu'offrent plusieurs modèles actuels est d'un niveau de complexité tel qu'un simple test empirique peut difficilement la valider. Il ne faut pas oublier l'hétérogénéité des

[3] Je tiens à remercier un des relecteurs anonymes pour cette idée.

facteurs à prendre en compte dans l'interprétation des résultats liés à la fréquence, et ce sur plusieurs niveaux de la représentation. A titre d'exemple, suivant Pierrehumbert (2006), l'effet de saturation et la non linéarité de la représentation du mot, l'effet de saillance résultant de l'encodage dans la mémoire, la dissociation de la perception et de la production, ou encore les effets liés aux réseaux complexes et la densité des connections. En d'autres termes une variable qui semble simple et direct ne l'est pas forcément.

# 4 Conclusions

Nous avons étudié les effets éventuels de la fréquence lexicale sur la dérhoticisation en Écosse et, au lieu d'un lien fort, nous avons trouvé une faible corrélation entre elles deux en comparant la rhoticité dans les quatre quartiles de la distribution totale des mots.

Pour une théorie phonologique sociocognitive, l'objectif est d'une part de modéliser l'usage de la langue dans la communauté, et d'autre part d'offrir une représentation mentale de la langue chez l'individu. Il n'est pas certain que les mêmes principes et processus puissent être utilisés dans les deux cas. Notre étude ne peut pas porter sur la diffusion des innovations avec un échantillon de dix-huit locuteurs même en faisant confiance à la puissance des tests statistiques. Il est aujourd'hui de plus en plus clair que l'étude du changement linguistique ne peut s'envisager que soit sur le niveau microscopique d'une étude ethnographique, soit sur un niveau beaucoup plus large avec des simulations computationnelles (Wagner & Abtahian, 2016). Ainsi, les différents facteurs phonologiques, sociophonétiques et psychologiques ne sont pas en compétition pour 'mener' un changement linguistique mais sont tous potentiellement concernés. Le défi est de faire des simulations où les agrégats de ces facteurs sont établis précisément.

Compte tenu de nos résultats, nous nous interrogeons sur la manière dont on doit aborder l'étude du changement linguistique : par la diffusion des formes innovantes que l'on peut observer dans la communauté ou par la représentation de la parole dans le système phonologique de l'individu ? Si l'approche sociocognitive a pour objectif d'expliciter les deux aspects, encore faut-il justifier que la même méthode de représentation est applicable aussi bien chez l'individu pendant et après la période d'acquisition qu'au sein de la population. La variation présente de manière inhérente et constante dans un groupe est-elle inscrite réellement dans la représentation mentale de chacun des locuteurs ? Ce qui semble être certain, c'est que le lien entre la diffusion des innovations phonétiques et l'évolution de leur schéma cognitif chez l'individu n'est pas direct.

# Références

CHEN M. & WANG W.S-Y. (1975). Sound change: Actuation and implementation. *Linguistics*. **51**. 255-8.

ERNESTUS M. & BAAYENS R. H. (2011). Corpora and exemplars in phonology. In GOLDSMITH J A, RIGGLE J & YU A C. (Eds.) *The handbook of phonological theory* (2nd ed.) Oxford : Wiley-Blackwell.

JAURIBERRY T. (2016). *Rhotiques et rhoticité en Ecosse : une étude sociophonétique de l'anglais écossais standard.* Thèse de doctorat. Université de Strasbourg.

JAURIBERRY T., SOCK R., HAMM A. & PUKLI M. (2012). Rhoticité et dérhoticisation en anglais écossais d'Ayrshire. Actes de *JEP*. 89-96.

JOHNTSON P. (1997). Regional variation. In JONES C. (Ed.) *The Edinburgh History of Scots*, Edinburgh : Edinburgh University Press, 433-513.

KILGARRIFF A. (1998). *BNC database and word frequency lists*. En-ligne : https://kilgarriff.co.uk/bnc-readme.html.

LAWSON E., STUART-SMITH J. & SCOBBIE J. (2018). The role of gesture delay in coda /r/ weakening: An articulatory, auditory and acoustic study. *The Journal of the Acoustical Society of America.* **143**, 1646. doi/10.1121/1.5027833.

LAWSON E., STUART-SMITH J. & SCOBBIE J. (2008). Articulatory insights into language variation and change: Preliminary findings from an ultrasound study of derhoticization in Scottish English. *University of Pennsylvania Working Papers in Linguistics.* **36**, 102-110.

NAGY N. & IRWIN P. (2010). Boston (r): Neighbo(r)s nea(r) and fa(r). *Language Variation and Change*, **22**, 241-278.

PIERREHUMBERT J. (2006). The next toolkit. *Journal of Phonetics* **34**(6), 516-530.

STUART SMITH J., LAWSON E. & SCOBBIE J. M. (2014). Derhoticisation in Scottish English: a sociophonetic journey. CELATA C. & CALAMAI S. (Eds.) *Advances in Sociophonetics*. John Benjamins, Amsterdam. http://eprints.gla.ac.uk/87460.

WAGNER S. E. & ABTAHIAN M. R. (2016). Social Networks and the Study of Language Variation and Change. In NEAL Z. P. (Ed.) *Handbook of Applied System Science*.

ZIPF G. K. (1929). *Relative frequency as a determinant of phonetic change*. Harvard Studies in classical philology, **XL**.

# L'opposition fortis / lenis des occlusives en fin de mot en anglais : liste de mots isolée lue par les apprenants francophones

Takeki Kamiyama[1, 2] Nadine Herry-Bénit[3] Ioana Trifu-Dejeu[3] Audrey Gros-Bonfiglioli[3]
(1)LeCSeL TransCrit EA1569, Paris 8-UPL 2 rue de la Liberté 93526 St Denis, France
(2) LPP UMR 7018, CNRS / Paris 3 USPC 19 rue des Bernardins 75005 Paris, France
(3) CREA EA370, Paris Nanterre-UPL 200 av. de la République 92001 Nanterre, France
takeki.kamiyama@univ-paris8.fr

## RESUME

L'opposition fortis / lenis en début de mot en anglais acquise par des locuteurs d'autres langues a été étudiée par de nombreux chercheurs, mais cette même opposition en fin de mot semble avoir été moins abordée dans des études empiriques, qui portent surtout sur les locuteurs de langues qui présentent une absence ou neutralisation de cette opposition en fin de mot (Smith et al., 2009 ; Skarnitzl & Šturm, 2016). Dans cette étude, 6 apprenantes francophones de l'anglais, qui connaissent une opposition de voisement en fin de mot dans leur première langue, ont prononcé 30 mots isolés. Il a été observé que la durée de la voyelle précédente, considérée comme indice primaire chez les anglophones natifs, a été produite avec une différence significative (test de la somme des rangs de Wilcoxon) par 5 apprenantes sur 6, mais certaines d'entre elles ont montré des réalisations phonétiques de voisement différentes des natifs (relâchement voisé pour lenis; fin de voisement plus tôt que les natifs).

## ABSTRACT

**Word-final fortis / lenis contrast in English plosives: lists of words in isolation read aloud by French-speaking learners**

The acquisition of the word-initial fortis / lenis contrast in English by speakers of other languages has been investigated by many researchers, but empirical studies on this contrast in word-final position seems to have been conducted less extensively, and especially on speakers of languages in which word-final voicing distinction is known to be absent or neutralized (Smith et al., 2009 ; Skarnitzl & Šturm, 2016). In this study, 6 French-speaking learners of English, who have voicing contrast word-finally in their first language, pronounced 30 words each in isolation. It was observed that the duration of the previous vowel, which is described as a primary cue in native speakers' speech, was produced with a significant difference (Wilcoxon rank sum test) by 5 learners out of 6, but some learners also showed cases of different voicing patterns for lenis than native speakers (voiced release for lenis; end of voicing earlier than natives).

MOTS-CLES : opposition fortis / lenis, occlusive, anglais, apprenant francophone, production, durée.

KEYWORDS: fortis / lenis contrast, plosive, English, French-speaking learner, production, duration.



10.21437/JEP.2018-25

# 1 Introduction

Le voisement des obstruentes de l'anglais en attaque (fortis ou voisées caractérisées par un VOT – temps d'établissement de voisement – long *vs* lenis ou non-voisées avec un VOT court et typiquement sans prévoisement), notamment en début de mot, est un des sujets les plus largement étudiés dans le domaine de l'acquisition des langues étrangères et secondes (Caramazza *et al*., 1973, Flege, 1987, entre autres), ou encore des troisièmes langues (Llama *et al*., 2010 ; Wrembel, 2014, entre autres). En revanche, cette même opposition en coda ou en fin de mot semble avoir attiré moins d'attention, et les études empiriques qui traitent ce phénomène portent essentiellement sur des locuteurs de langues présentant une absence de voisement ou d'aspiration en coda (Cunningham, 2009 pour le vietnamien) ou une neutralisation au moins partielle de voisement en fin de mot (Smith *et al*., 2009 pour l'allemand, Skarnitzl & Šturm, 2016 pour le tchèque).

Dans le premier cas, Cunningham (2009) a mesuré les durées segmentales dans les mots « bead » /bid/ - « beat » /bit/ - « bid » /bɪd/ - « bit » /bɪt/ prononcés dans une phrase cadre par 3 locutrices vietnamiennes. La voyelle est plus brève dans « beat » que dans « bead » chez les apprenantes, mais la différence est moins marquée que chez l'anglophone natif étudié (décrit comme locuteur du RP, *Received Pronunciation*). En revanche, celle de « bit » est plus longue que dans « bid », ce qui montre une tendance inverse par rapport à celle observée chez les anglophones natifs.

Dans le dernier cas, les obstruentes voisées sont communément décrites comme dévoisées en fin de mot, mais il a été montré que l'opposition n'est pas toujours totalement perdue (neutralisation partielle : Warner et al. 2004 pour le néerlandais et Smith et al., 2009 pour l'allemand). Cependant, la distinction de voisement en anglais n'est pas acquise avec facilité. Smith et al., 2009 a montré que les 13 locuteurs germanophones natifs étudiés, qui ont lu 26 mots cible en allemand et 30 mots cible en anglais dans des phrases porteuses, ont typiquement produit moins d'indices acoustiques ou des indices moins robustes pour le voisement de fin de mot que les anglophones natifs, même s'ils ont présenté plus d'indications de distinction de voisement quand ils ont lu des mots en anglais que les mots phonologiquement similaires en allemand. Dans Skarnitzl & Šturm (2016), 10 apprenants tchécophones de l'anglais ont lu 32 phrases contenant 16 paires minimales monosyllabiques (« rich » /rɪtʃ/ - « ridge » /rɪdʒ/, « stack » /stæk/ - « stag » /stæg/, etc.). Leurs résultats indiquent une absence de différence significative de durées vocaliques, quelle que soit la catégorie de consonne finale que les voyelles précèdent.

À la différence de ces langues, le français présente une opposition de voisement dans toutes les positions, mais la réalisation phonétique diffère de celle de l'anglais. Les obstruentes voisées sont en général caractérisées comme entièrement voisées et les non-voisées sont prononcées avec un VOT court dans de nombreuses variétés du français. En anglais, les obstruentes voisées, ou lenis, en fin de mots sont souvent partiellement dévoisées. Quant à elles, les non-voisées, ou fortis, sont caractérisées par un raccourcissement de la voyelle précédente (*pre-fortis clipping*). Cette tendance est observée dans de nombreuses langues, y compris le français, mais la différence est particulièrement marquée en anglais (Chen, 1970). Selon Kohler (1994), la durée de la voyelle devant les obstruentes fortis et lenis a été phonologisée en anglais. Par ailleurs, le ratio entre la voyelle et la consonne est également considéré pour cette opposition en fin de syllabe (Massaro & Cohen, 1983, Port & Dalby, 1982, pour les occlusives à l'intervocalique, du moins).

En nous fondant sur les études antérieures, nous pouvons émettre l'hypothèse que la production de l'opposition de voisement en fin de mot en anglais n'est pas nécessairement facilitée par l'existence de l'opposition de voisement en français dans cette position, suite à des réalisations phonétiques différentes entre les deux langues. Afin d'étudier la production des apprenants francophones de l'anglais langue étrangère, une étude de production a été effectuée.

# 2 Méthode

Dans la présente étude, 30 mots (15 paires minimales) monosyllabiques CVC qui se terminent par une occlusive fortis ou lenis, montrés dans la Table 1, ont été retenus des listes de lectures du protocole ICE-IPAC (*Interphonology of Contemporary English* – InterPhonologie de l'Anglais Contemporain : Herry-Bénit et al., en préparation). Huit d'entre eux (4 paires minimales) sont également inclus dans le protocole PAC (Phonologie de l'Anglais Contemporain : Brulard et al., 2015), ce qui permet une comparaison avec la production des locuteurs natifs. Tous les mots cible ont été prononcés avec d'autres mots dans deux listes, composées chacune de 72 et de 92 mots isolés. Chaque mot a été précédé d'un numéro d'identification et d'une pause (ex. « *fourteen* (pause) *lab* »). Les mots ont été présentés un par un sur un écran d'ordinateur.

| Mots avec une occlusive fortis | | | Mots avec une occlusive lenis | | |
|---|---|---|---|---|---|
| lap /læp/ (A6) | bat /bæt/ (A32) | sack /sæk/ (A61) | lab /læb/ (A14) | bad /bæd/ (B51) | sag /sæg/ (A34) |
| | fat /fæt/ (B67) | lack /læk/ (B10) | | fad /fæd/ (B39) | lag /læg/ (A44) |
| | pat /pæt/ (A4) | | | pad /pæd/ (A23) | |
| | lat /læt/ (B20) | | | lad /læd/ (A28) | |
| | beat /bit/ (B14) | | | bead /bid/ (B84) | |
| dip /dɪp/ (B17) | bit /bɪt/ (A10) | Dick /dɪk/ (A12) | dib /dɪb/ (B37) | bid /bɪd/ (B63) | dig /dɪg/ (A49) |
| | dit /dɪt/ (B44) | | | did /dɪd/ (B6) | |
| cop /kɑp/ (B56) | cot /kɑt/ (B9) | cock /kɑk/ (A40) | cob /kɑb/ (B32) | cod /kɑd/ (B47) | cog /kɑg/ (A21) |

TABLE 1 : Les 30 mots retenus pour l'analyse. La transcription phonémique correspond à *General American*. « A » (première) et « B » (deuxième) correspondent aux deux listes de mots à lire dans le protocole ICE-IPAC, et le numéro, l'ordre dans chacune des listes. Les mots représentés en gris forment des paires minimales incluses également dans les listes PAC (lues par 2 locuteurs anglophones natifs).

Les enregistrements de 6 locutrices francophones natives (FR1 – FR6) ont été utilisés pour les analyses. Les locutrices étaient âgées de 20 à 30 ans, originaires de Paris et de Lyon. Le niveau de compétences générales en anglais variait de A1 à B1 du CECRL (Cadre Européen Commun de Référence pour les Langues). L'enregistrement à été effectué dans le studio d'enregistrement de l'Université Paris 8 (Saint-Denis). Les données ont été sauvegardées au taux d'échantillonnage de 44100 Hz, 16 bits. Par ailleurs, les enregistrements de 2 locuteurs natifs de l'anglais américain qui ont effectué la tâche de lecture de mots du protocole PAC (AM1, AM2) ont été soumis à des analyses.

Afin de permettre des mesures de durée, les événements acoustiques suivants ont été marqués pour chaque mot cible :

1) le début de la voyelle (apparition du deuxième formant – F2) ;
2) la fin de la voyelle (disparition du F2) ;
3) la fin du voisement (disparition complète de la barre de voisement) ;
4) le relâchement de la consonne de coda, si audible (4 cas de relâchement non audible observés sur 196, dont 2 pour le /t/ chez AM1, 2 pour le /b/ chez FR5 et FR6) ;
5) la fin de la consonne de coda (fin du bruit).

La fin du voisement (3) peut se trouver avant le relâchement de la consonne de coda, ou après, si cette consonne est entièrement voisée, et accompagnée éventuellement d'un vocoïde. Quand le voisement s'arrête légèrement avant le relâchement de la consonne de coda et reprend après le relâchement pour un vocoïde (5 cas observés dans la production des 98 occlusives lenis analysées, et une seule occurrence d'une fortis accompagnée d'un vocoïde chez une apprenante francophone) la durée de voisement pour le vocoïde a également été marquée. Dans l'exemple illustré dans la Figure 1, le voisement s'arrête (v3) avant le relâchement de la consonne de coda (c1), mais le voisement ne reprend pas après le relâchement.

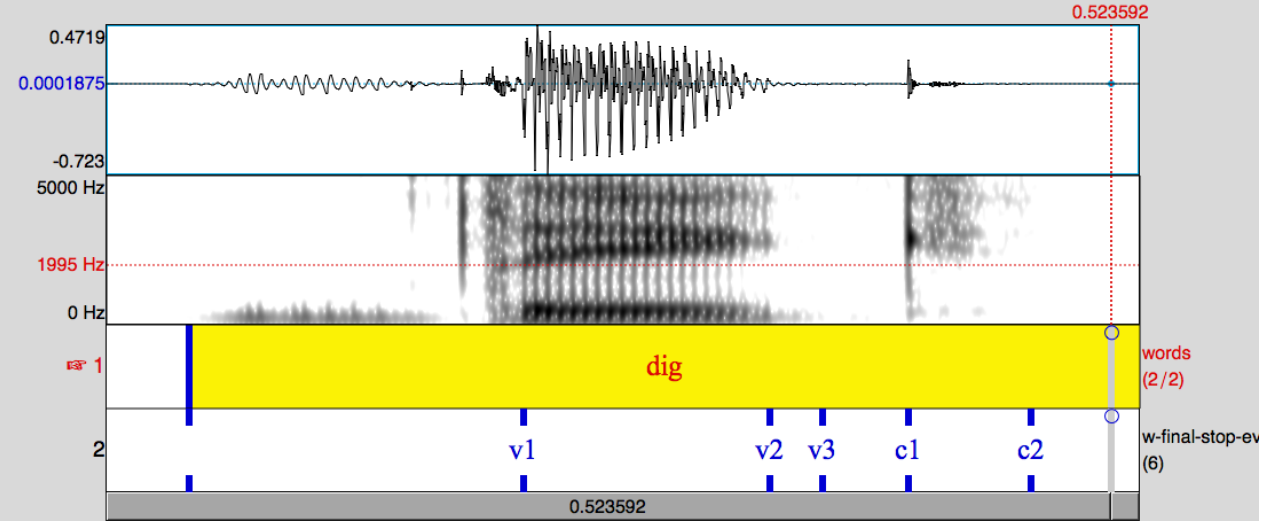



FIGURE 1 : Exemple de mesures de durée (« dig » /dɪg/ prononcé par l'apprenante FR6). v1 : début de la voyelle ; v2 : fin de la voyelle ; v3 : fin du voisement ; c1 : relâchement de la consonne en coda ; c2 : fin de la consonne (du bruit).

En se fondant sur ces marques, les durées suivantes ont été calculées :

1) durée brute de la voyelle (distance entre v1 et v2 dans la Figure 1) ;
2) durée relative de la voyelle, par rapport à la distance entre le début de la voyelle et le relâchement de la consonne de coda, si audible (distance entre v1 et v2 / distance entre v1 et c1) ;
3) durée totale brute du voisement (distance entre v1 et v3 + durée du voisement après le relâchement de la consonne, si applicable) ;
4) durée totale relative du voisement, par rapport à la distance entre le début de la voyelle et le relâchement de la consonne de coda, si audible ((distance entre v1 et v3 + durée du voisement après le relâchement de la consonne, si applicable) / distance entre v1 et c1) ;
5) durée du relâchement de la consonne de coda (distance entre c1 et c2).

# 3 Résultats

Trois d'entre ces mesures de durée seront présentés ici : 1) la durée brute de la voyelle ; 2) la durée de la voyelle relative à la distance entre le début de la voyelle et le relâchement de la consonne de coda ; 3) la durée totale du voisement relative à la distance utilisée dans 2).

## 3.1 Durée brute de la voyelle

La durée brute de la voyelle est présentée dans la Figure 2, qui indique la distribution des valeurs mesurées sur l'ensemble des 30 mots prononcés par les apprenantes (à gauche) et celle des 8 mots produits également par les 2 locuteurs natifs permettant une comparaison avec les valeurs des natifs (à droite). Parmi les 6 apprenantes francophones, 5 (FR1, 2, 4, 5 et 6) ont produit la voyelle avec une durée significativement plus longue devant une occlusive lenis en coda que devant une occlusive fortis (respectivement p < ,001 ; < ,05 ; < ,01 ; < ,001 ; < ,001, test de la somme des rangs de Wilcoxon : la distribution des mesures ne peut pas être considérée normale pour toutes les locutrices). Une comparaison avec les locuteurs natifs AM1 et 2 (Figure 2, à droite) montre que les deux catégories se chevauchent chez les apprenantes (sauf FR5), à la différence de AM 1 et 2, qui présentent la tendance connue dans la littérature.

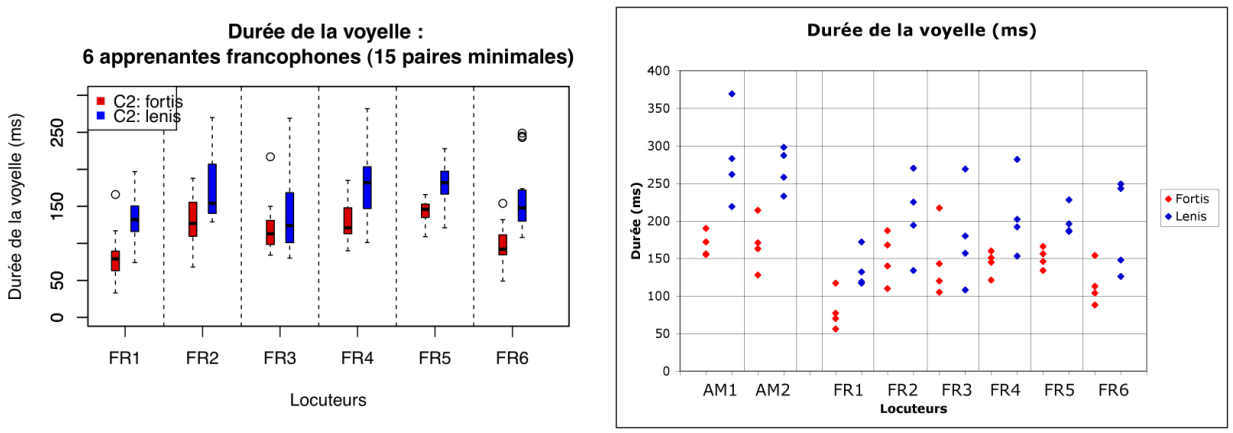
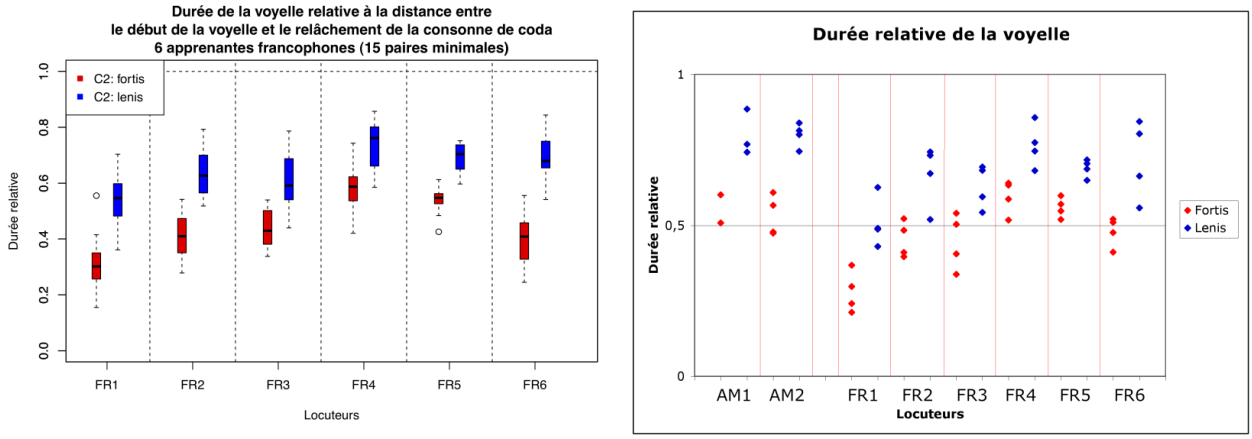



FIGURE 2 : Durée brute (en ms) de la voyelle dans des mots CVC : (à gauche) production des 30 mots par 6 apprenantes FR1 – FR 6 ; (à droite) 8 mots (4 paires minimales) prononcés par 2 anglophones natifs américains AM1 – AM2 et 6 apprenantes francophones FR1 – FR 6.

## 3.2 Durée relative de la voyelle

La Figure 3 montre la durée de la voyelle relative, qui tient compte de la durée de l'occlusion de l'occlusive de coda, car la valeur de référence correspond à la distance entre le début de la voyelle et le relâchement de la consonne de coda. Les apprenantes francophones montrent toutes une durée relative de la voyelle significativement plus longue pour les lenis que pour les fortis (p < ,001, test t de Student). En revanche, cette différence s'avère moins robuste que chez AM 1 et 2 quand on analyse les mots communs avec les productions des natifs. On observe des chevauchements pour FR2 et 3, et l'écart entre les deux catégories, fortis et lenis, est moins important chez les

apprenantes que chez les natifs. On remarque également que la distribution des valeurs des FR2-6 (sauf FR1) pour les fortis est plus congruente avec celle de AM1-2 que pour les valeurs des lenis, ce qui laisse de nombreux cas de production de lenis (comme cible) chez les apprenantes qui se situent dans la zone des valeurs des fortis chez les natifs. Cela indique que, dans ces cas-là, les apprenantes produisent, pour une cible lenis, une occlusion relativement longue, comparable à celle des fortis chez les natifs.




FIGURES 3 : Durée relative de la voyelle : (à gauche) 30 mots, sauf pour FR5 et 6 (29 mots, suite à un /b/ non relâché), prononcés par les apprenantes ; (à droite) 2 locuteurs anglophones américains natifs (AM1, AM2) et 6 apprenantes francophones (FR1 – FR6). 8 mots (4 paires minimales) par locuteur, sauf AM1 (6 mots : les 2 mots prononcés sans relâchement audible du /t/ final exclus).

## 3.3 Fin du voisement

La Figure 4 montre la durée totale du voisement relative à la distance entre le début de la voyelle et le relâchement de la consonne de coda, ce qui donne une indication du moment de l'arrêt du voisement par rapport au moment du relâchement : les valeurs proches de 1 correspondent à un arrêt de voisement dans une zone proche du relâchement de l'occlusive.

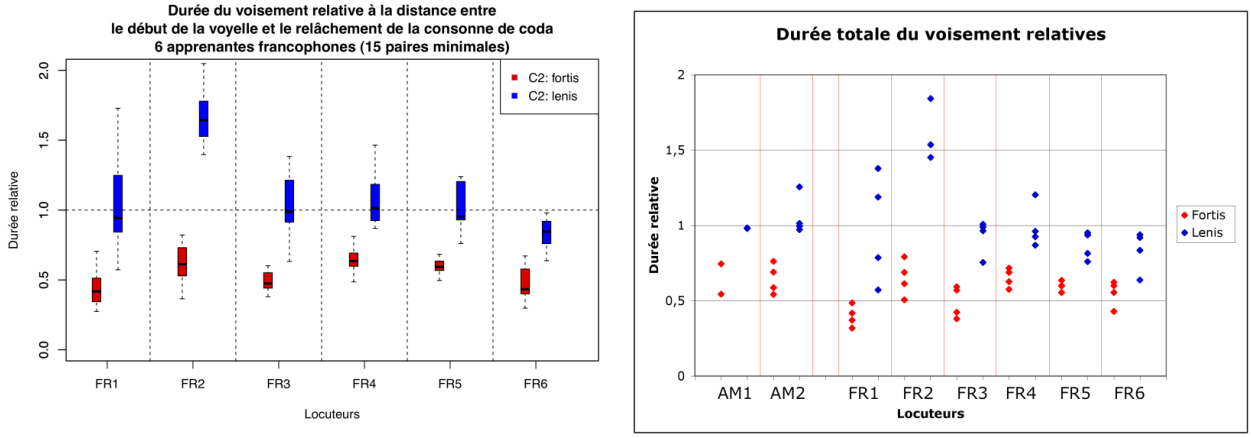


FIGURE 4 : Durée totale relative du voisement (entre le début de la voyelle et la fin du voisement + la durée du voisement du vocoïde après le relâchement de la consonne de coda, si applicable) : (à gauche) les apprenantes francophones. 30 mots par locuteur, sauf pour FR5 et 6 (29 mots, suite à un /b/ non relâché) ; (à droite) 2 locuteurs anglophones américains natifs AM1 – AM2 et 6 apprenantes francophones FR1 – FR6. 8 mots (4 paires minimales) par locuteur, sauf AM1 (6 mots : les 2 mots prononcés sans relâchement audible du /t/ final exclus).

Chez toutes les apprenantes, la durée relative du voisement était significativement plus longue pour les lenis que pour les fortis ($p < ,001$, test de la somme des rangs de Wilcoxon). FR1, 3, 4 et 5 présentent des valeurs qui varient autour de 1 pour les lenis, mais avec celles qui s'en éloignent également dans les deux sens. Chez FR2, les valeurs sont largement supérieures à 1 pour les lenis, ce qui signifie la présence systématique d'un relâchement voisé ou d'un vocoïde. FR6 montre des valeurs proches de, mais ne dépassant pas 1 pour les lenis. Cette tendance est comparable à celle des locuteurs natifs AM1 et 2 (mise à part une occurrence chez AM2), qui marquent la fin du voisement dans une zone temporelle qui correspond à peu près au moment du relâchement des occlusives lenis. En revanche, FR6 présente également des occurrences avec des valeurs inférieures. Tout comme la durée relative de la voyelle, certaines valeurs de la production de lenis (chez FR1, 3, 5 et 6) se trouvent dans la zone des valeurs des fortis chez les natifs, indiquant un arrêt de voisement précoce comparable à celui de la production des fortis chez les natifs.

# 4 Discussion et conclusion

Les résultats présentés ci-dessus montrent que les apprenantes francophones ont tendance à produire les deux catégories d'occlusives, fortis et lenis, à la fin des mots étudiés, de manière distinctive, même si les différences de tous les indices acoustiques mesurés, la durée de la voyelle, de l'occlusion de la consonne et du voisement, ne sont pas aussi robustes que les locuteurs natifs. Certaines mesures semblent révéler l'état de l'interlangue dans laquelle les apprenants tâtonnent à la recherche de la cible. Par exemple, la fin du voisement (durée relative du voisement) montre que certaines (notamment FR2) produisent des valeurs proches de la tendance ouvrante du français (avec un vocoïde après le relâchement), alors que d'autres varient entre des productions à la L1 (valeur supérieure à 1) et des formes d'hyper-correction, dans lesquelles le voisement s'arrête relativement tôt comme pour les fortis de la langue cible.

Ce dernier type de production pourrait être perçu comme fortis par les auditeurs natifs. L'évaluation perceptive par des locuteurs natifs, qui n'a pas été effectuée dans la présente étude, sera indispensable, même si les paramètres acoustiques pertinents dans le mécanisme de la perception de fortis / lenis semblent relativement complexes (Hillenbrand et al. 1984). Comment seront perçues des occlusives lenis produites par des apprenantes avec une voyelle suffisamment longue mais avec un arrêt de voisement précoce par rapport au relâchement (comme l'exemple montré dans la Figure 1) ?

Cette étude porte sur la production des apprenants francophones, mais leur perception de fortis-lenis, notamment des lenis, en fin de mot ainsi que la représentation de cette opposition sera un facteur non négligeable. Il est bien possible que la production soit liée à la représentation des apprenants (Tilsen & Cohn, 2016, pour la représentation du nombre de syllabes dans des mots qui se terminent par /l/ et la durée dans les productions de locuteurs natifs de l'anglais). Les apprenants pourraient ainsi identifier comme fortis des lenis en fin de mot produites par des natifs sans relâchement voisé à la française.

Parmi les mesures effectuées dans cette étude, la durée vocalique, qui n'est pas distinctive en français au niveau segmental, mais est utilisée essentiellement au niveau prosodique pour marquer les frontières, serait difficile à utiliser au niveau segmental pour les francophones (Dupoux et al. 1999 pour la difficulté de perception). Il sera intéressant d'examiner si une autre forme de distinction de durée vocalique (voyelles phonologiquement brèves et longues), pour ceux qui l'ont dans la L1, faciliterait l'acquisition de l'utilisation de la durée pour distinguer les fortis et les lenis en fin de mot en anglais.

Le relâchement de l'occlusive de coda est un facteur qui n'a pas été traité dans le présent article, mais la durée ainsi que la structure spectrale devra être également prise en compte.

Concernant les apprenants d'autres langues maternelles, il a été observé que des apprenants natifs du norvégien présentent des cas de dévoisement des occlusives lenis en fin de mots, où le voisement s'arrête bien avant le relâchement (Andreassen et al., 2015). Notons que le norvégien (comme les autres langues germaniques du nord) ne fait pas partie des langues présentant une neutralisation de l'opposition de voisement des occlusives en fin de mot, sauf qu'il n'y a pas d'opposition de voisement pour les fricatives, quelle que soit la position. Il conviendra de comparer les cas de l'acquisition de l'anglais langue étrangère et seconde par les apprenants de langues typologiquement diverses concernant l'opposition de voisement en fin de mot : langues sans opposition (vietnamien, cantonais, …) ; langue sans obstruente en coda (japonais, chinois mandarin, …) ; langue avec neutralisation au moins partielle (allemand, néerlandais, tchèque, polonais, catalan, …) etc.

# Remerciements

Les auteurs remercient les deux relecteurs anonymes pour leurs conseils constructifs, ainsi que les membres du projet PAC et les participants aux workshops PAC des années précédentes pour leur commentaires pendant l'étape de démarrage de ce travail de recherche.

# Références

ANDREASSEN, H. N., HERRY-BÉNIT, N., KAMIYAMA, T., LACOSTE, V. (2015). The ICE-IPAC project: testing the protocol on Norwegian and French learners of English. Poster présenté au Workshop on Phonetic Learner Corpora, réunion satellite de l'ICPhS 2015, Glasgow, Écosse.

BOERSMA P., WEENINK D. (2015). *Praat: doing phonetics by computer* [logiciel], Version 5.4.14 téléchargé en août 2015 depuis http://www.praat.org/.

BRULARD, I, CARR, P., DURAND, J. (2015). *La Prononciation de l'anglais contemporain dans le monde : variation et structure*. Toulouse : Presses Universitaires du Midi.

CARAMAZZA, A., YENI-KOMSHIAN, G., ZURIF, E., CARBONE, E. (1973). The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals. *Journal of the Acoustical Society of America* 54, 421-428.

CHEN, M. (1970). Vowel length variation as a function of the voicing of the consonant environment. *Phonetica* 22, 129–159.

CUNNINGHAM, U. (2009). Quality, quantity and intelligibility of vowels in Vietnamese accented English. Waniek-Klimczak, E. (éd), *Issues in Accents of English II: Variability and Norm*. Newcastle: Cambridge Scholars Publishing Ltd., 1-15.

DUPOUX, E., KAKEHI, K., HIROSE, Y., PALLIER, C., MEHLER, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance* 25(6), 1568-1578.

FLEGE, J. E. (1987). The production of "new" and "similar" phones in a foreign language: evidence for the effect of equivalence classification. *Journal of Phonetics* 15, 47-65.

HERRY-BÉNIT, N., TENNANT, J., KAMIYAMA, T. (IN PREPARATION). ICE-IPAC (Interphonology of Contemporary English) project: methodological issues.

HILLENBRAND, J., INGRISANO, D. R., SMITH, B. L., FLEGE, J. E. (1984). Perception of the voiced–voiceless contrast in syllable-final stops. *Journal of the Acoustical Society of America* 76, 18–26.

KOHLER, K. J. 1984. Phonetic explanation in phonology: the feature fortis/lenis. *Phonetica* 41, 150–174.

LLAMA, R., CARDOSO, W., COLLINS, L. (2010). The influence of language distance and language status on the acquisition of L3 phonology. *International Journal of Multilingualism* 7(1), 39-57.

MASSARO, D.W., COHEN, M.M. (1983) Consonant/vowel ratio: An improbable cue in speech. *Perception & Psychophysics* 33, 502-505.

PORT, R. F., DALBY, J. (1982). Consonant/vowel ratio as a cue for voicing in English. *Perception & Psychophysics* 32, 141-152.

SKARNITZL, R., ŠTURM, P. (2016). Pre-Fortis Shortening in Czech English: A Production and Reaction-Time Study. Research in Language 14(1), 1-14.

SMITH, B. L., HAYES-HARB, R., BRUSS, M. HARKER A. (2009). Production and perception of voicing and devoicing in similar German and English word pairs by native speakers of German. *Journal of Phonetics* 37, 257–275.

TILSEN, S., COHN, A. C. (2016). Shared Representations Underlie Metaphonological Judgments and Speech Motor Control. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7(1), 1-33.

WARNER, N., JONGMAN, A., SERENO, J., KEMPS, R. (2004). Incomplete neutralization and other sub-phonemic durational differences in production and perception: evidence from Dutch. *Journal of Phonetics* 32, 251–276.

WREMBEL M. (2014). VOT Patterns in the Acquisition of Third Language Phonology. *Concordia Working Papers in Applied Linguistics* 5, 751-771.

# Perception de la parole et oscillations cérébrales chez les enfants neurotypiques et dysphasiques

Hélène Guiraud[1], Ana-Sofia Hincapié[3], Karim Jerbi[2] and Véronique Boulenger[1]
(1) Laboratoire Dynamique Du Langage, CNRS/Université Lyon 2 UMR5596, 14 avenue Berthelot, 69007 Lyon, France
(2) Département de Psychologie, Université de Montréal, 90 avenue Vincent d'Indy, Montréal QC H2V2S9, Québec, Canada
(3) Pontificia Universidad Catolica de Chile, Av Libertador Bernardo O'Higgins, Santiago de Chile, Chile
guiraudh@gmail.com, ana.hincapie@gmail.com, karim.jerbi@umontreal.ca, veronique.boulenger@cnrs.fr

## RESUME

L'hypothèse "*prosodic phrasing*" (Cumming et al., 2015) suggère que les enfants dysphasiques présentent des difficultés d'extraction des informations rythmiques de basse fréquence de la parole, entravant la segmentation syllabique et conduisant à des déficits phonologiques et morphosyntaxiques. Nous avons testé cette hypothèse en mesurant, en magnétoencéphalographie, la synchronisation entre les oscillations cérébrales et l'enveloppe temporelle du signal de parole chez des enfants dysphasiques et des enfants neurotypiques lors de l'écoute de phrases produites naturellement à un débit normal ou rapide. Nos résultats dans la bande de fréquence thêta (4-7 Hz) montrent une synchronisation plus faible chez les enfants dysphasiques, comparés aux enfants neurotypiques, 1) dans les régions auditives droites pour la parole à débit normal et 2) dans les régions (pré)motrices gauches pour la parole à débit rapide. Notre étude fournit les premiers éléments à notre connaissance en faveur d'un alignement cortical atypique sur le rythme syllabique dans la dysphasie.

## ABSTRACT

According to the "prosodic phrasing" hypothesis (Cumming et al., 2015), children with Developmental Language Disorder (DLD) show difficulty extracting low-frequency rhythmic information from the speech signal, hindering syllabic segmentation and leading to phonological and morpho-syntactic impairments. We tested this hypothesis by measuring, using magnetoencephalography, the synchronization between cortical oscillations and speech amplitude envelope in children with DLD paired to typically-developing children when listening to sentences naturally produced at a normal or rapid rate. Our results in the theta frequency band (4-7 Hz) show reduced brain-to-speech coupling in children with DLD, as compared with typically-developing children, 1) in the right auditory cortex at normal rate and 2) left (pre)motor regions at fast rate. To our knowledge, this study brings the first piece of evidence for atypical cortical alignment to speech syllabic rhythm in children with DLD.

MOTS-CLÉS : parole, oscillations cérébrales, développement typique, dysphasie, débit
KEYWORDS: speech, brain oscillations, typical development, Developmental Language Disorder, speech rate


10.21437/JEP.2018-26

# 1 Introduction

La dysphasie est un trouble neurodéveloppemental sévère et durable de l'élaboration du langage oral qui ne peut être attribué à un trouble auditif, cognitif ou neurologique, un trouble du spectre autistique ou une carence affective ou éducative (Maillart, Schelstraete, 2012). La dysphasie peut affecter l'ensemble des composantes langagières de l'expression et/ou de la compréhension avec une sévérité variable ; les difficultés sont néanmoins le plus souvent décrites sur le plan phonologique et morphosyntaxique (i.e. dysphasie phonologico-syntaxique). Pour rendre compte de ces déficits, Cumming et collaborateurs (2015) ont avancé l'hypothèse de « *prosodic phrasing* » selon laquelle les enfants dysphasiques présenteraient une sensibilité réduite aux informations rythmiques contenues dans l'enveloppe temporelle de la parole, dont les modulations d'amplitude dominent dans les basses fréquences (4-7 Hz, thêta) et reflètent le débit syllabique du locuteur (Ghitza, Greenberg, 2009). Cette insensibilité entraverait alors la segmentation du signal en syllabes et en mots, et par conséquent le traitement de la structure phonologique et morphosyntaxique de la parole. Des études ont révélé que deux indices suprasegmentaux importants pour le traitement du rythme et de l'accentuation de la parole, le *rise time* (temps de montée des modulations successives de l'enveloppe d'amplitude) et la durée du signal, étaient particulièrement difficiles à traiter pour les enfants dysphasiques ; ces difficultés étaient en outre corrélées à leurs performances phonologiques (Corriveau et al., 2007; Cumming et al., 2015). Nous avons par ailleurs montré des performances réduites chez les enfants dysphasiques, en regard d'enfants au développement typique (dits neurotypiques), pour comprendre la parole accélérée naturellement ou artificiellement, suggérant un déficit de traitement du rythme de la parole dès lors qu'il est accéléré (Guiraud et al., sous presse).

Soulignant la correspondance étroite entre le rythme de la parole et les rythmes intrinsèques du cerveau, les modèles neurocognitifs actuels suggèrent que les oscillations corticales jouent un rôle majeur dans le décodage du signal verbal (Ghitza, 2011 ; Giraud, Poeppel, 2012). En particulier, les oscillations du cortex auditif dans la bande de fréquence thêta (4-7 Hz) seraient capables de se synchroniser sur l'enveloppe temporelle de la parole, permettant au cerveau de l'auditeur de segmenter le flux continu en unités pertinentes, les syllabes, et d'ainsi faciliter le traitement linguistique. Il a été suggéré que les déficits de traitement rythmique dans les troubles développementaux du langage puissent résulter d'une synchronisation atypique des oscillations corticales sur le signal de parole, en particulier dans les basses fréquences (Giraud, Poeppel, 2012 ; Goswami, 2011). Bien que l'hypothèse d'un dysfonctionnement cortical oscillatoire dans la dysphasie soit particulièrement séduisante, elle n'a jamais fait l'objet d'investigations à l'heure actuelle et à notre connaissance. Des travaux ont néanmoins révélé un fonctionnement oscillatoire atypique dans la dyslexie, caractérisé par un trouble de la synchronisation des oscillations corticales sur le signal de parole dans les basses et hautes fréquences (*e.g.*, Molinaro et al., 2016).

Dans cette étude, nous avons examiné, à l'aide d'une technique d'imagerie cérébrale non invasive, la magnétoencéphalographie (MEG), la dynamique des oscillations cérébrales lors de la perception de parole chez des enfants dysphasiques et des enfants neurotypiques. Nous nous sommes intéressés au cas de la parole produite naturellement à un débit normal ou rapide, afin de déterminer dans quelle mesure les oscillations corticales étaient capables de « suivre » le rythme syllabique. Notre hypothèse était que chez les enfants neurotypiques, les régions auditives devraient synchroniser leur activité oscillatoire thêta (4-7 Hz) sur la parole dans les deux conditions de débit, avec l'implication

éventuelle de régions supplémentaires pour traiter la parole rapide, plus difficile à comprendre. Chez les enfants dysphasiques, en accord avec le modèle de Cumming et al. (2015) et notre étude comportementale (Guiraud et al., sous presse), nous nous attendions à une synchronisation thêta atypique (i.e. couplage réduit et/ou réseaux corticaux différents) entre rythmes cérébraux et linguistiques, comparés aux enfants neurotypiques, pour les deux conditions de débit.

# 2 Matériel et méthode

## 2.1 Participants

Onze enfants présentant une dysphasie phonologico-syntaxique (DYS, âge moyen 10.29 ans, Ecart-Type OU ET 1.54), appariés en âge et sexe à 11 enfants neurotypiques (NT, âge moyen 10.83 ans, ET 1.65), ont participé à l'étude. Tous les enfants étaient âgés de 8 à 13 ans, de langue maternelle française (non bilingues), droitiers et sans trouble de l'audition (vérification par audiométrie tonale). Les enfants DYS respectaient différents critères d'inclusion (QI non-verbal > 70 et troubles prédominants sur le versant expressif avec une compréhension préservée) et d'exclusion (déficit de l'attention/hyperactivité, trouble du spectre autistique, retard intellectuel). Avant l'expérience, les capacités verbales (phonologie, morphosyntaxe, compréhension, vocabulaire) de tous les enfants ont été évaluées grâce à différents sous-tests de la BALE (Batterie Analytique du Langage Écrit) et de l'ELO (Évaluation du Langage Oral). Leurs capacités non-verbales ont été examinées à l'aide des matrices progressives colorées de Raven et de l'empan endroit et envers de chiffres. Enfin, le débit de parole des enfants a été mesuré dans une tâche narrative de description d'images ("*Frog where are you*?") : il était de 1.81 syll/s (ET 0.43) chez les DYS et de 2.45 syll/s (ET 0.70) chez les NT ($p$ = .02). Le protocole était conforme à la déclaration d'Helsinki et a reçu un avis favorable du Comité de Protection des Personnes (ID RCB: 2012-A00857-36). Tous les enfants et leurs parents ont signé un formulaire de consentement éclairé avant l'expérience.

## 2.2 Stimuli

Trois cents phrases (7-9 mots) de structure syntaxique identique (*e.g.,* "Le public applaudit le joueur pour sa victoire") ont été enregistrées (44.1 kHz, mono, 16 bits) par un locuteur français de sexe masculin dans une salle insonorisée à l'aide du logiciel *ROCme*!. Chaque phrase a été enregistrée à un débit normal (moyenne 6.61 syll/s, ET 0.47) puis rapide (moyenne 9.03 syll/s, ET 0.56). Un filtre passe-haut de 80 Hz ainsi qu'un fondu d'entrée et de sortie sur l'enveloppe d'amplitude ont été appliqués sur la totalité des fichiers sons avec le logiciel Praat; le pic d'intensité a également été normalisé. Les 600 phrases (2×300) ont été réparties dans deux listes expérimentales de 300 stimuli (150 à débit normal, 150 à débit rapide). La fréquence d'occurrence dans la langue des mots finaux des phrases, ainsi que leurs nombres de phonèmes et de voisins phonologiques, ne différaient pas significativement entre les listes. Chaque phrase apparaissait dans chaque condition de débit à travers tous les participants, mais une seule fois par liste. Quarante phrases « *fillers* » (20 de chaque débit, non analysées), similaires aux phrases cibles mais à la fin desquelles un son était ajouté (http://www.sound-fishing.net/), ont été créées et ajoutées à chaque liste expérimentale pour un total

de 340 stimuli par liste. Au sein de chaque liste, les stimuli étaient pseudo-randomisés (pas plus de 3 phrases consécutives de la même condition). Afin d'analyser le couplage entre le signal de parole et l'activité cérébrale oscillatoire, l'enveloppe d'amplitude de chaque phrase a été extraite selon la méthode de Peelle et al. (2013 ; signal rectifié et filtré à 30 Hz avec un filtre passe-bas). Cette enveloppe a constitué le signal acoustique utilisé dans les analyses présentées en 2.4.

## 2.3 Procédure

Les enfants étaient confortablement allongés dans une chambre d'enregistrement insonorisée et blindée, la tête positionnée dans le casque MEG (système « tête entière » à 275 canaux, OMEGA MEG CTF 275 ; échantillonnage à 1200 Hz). Ils étaient équipés d'écouteurs intra-auriculaires et un écran était placé à 40 cm devant eux. Avant de débuter l'enregistrement MEG, un seuil auditif était réalisé pour chaque oreille afin de vérifier que tous les enfants entendaient les stimuli à la même intensité (détection d'un son pur de 400 à 3000 Hz). L'intensité était ensuite ajustée de sorte que les stimuli verbaux soient présentés en stéréo à 50 dB SL. Au cours de l'enregistrement, les enfants avaient pour consigne d'écouter attentivement les phrases (en fixant une croix sur l'écran) et d'appuyer sur un bouton avec leur index gauche lorsqu'ils entendaient un stimulus *filler*. Cinq phrases d'entrainement étaient proposées avant le début de la tâche. L'expérience a été réalisée avec le logiciel Présentation (*Neurobehavioral Systems*).

## 2.4 Analyses MEG

Les analyses ont été réalisées avec Matlab (Mathworks Inc., MA, USA) et la *toolbox* Fieldtrip. Après avoir ré-échantillonné les données à 300 Hz et supprimé les artéfacts musculaires, oculaires et cardiaques, des analyses de sources ont été réalisées. Pour cela, un modèle (*template*) de cerveau enfant (pour la tranche d'âge 7.5-13.5 ans), aligné sur les points fiduciaux (nasion et points pré-auriculaires) des participants de notre expérience, a été utilisé afin de créer un modèle de tête individuel pour chaque enfant. Ces modèles individuels ont ensuite été transposés sur un *template* adulte afin d'effectuer les analyses de sources dans un système de coordonnées commun (*Montreal Neurological Institute* ou MNI). Ceci nous a permis d'utiliser un atlas anatomique adulte pour les analyses statistiques basées sur des régions d'intérêt (ROIs). Nous avons mesuré la cohérence entre le signal MEG et l'enveloppe d'amplitude du signal de parole (i.e. cohérence cortico-acoustique) ainsi que les modulations de puissance au niveau des sources corticales à l'aide d'un filtre spatial (DICS). La cohérence correspond ici à la relation linéaire entre l'amplitude de deux signaux dans le domaine fréquentiel. Pour chaque essai, nous avons défini une période active (0.2 à 1 s après le début de la phrase) et une période de référence (intervalle inter-stimulus, -0.2 à -1 s avant la phrase). La cohérence cortico-acoustique a été calculée pour toutes les périodes actives et de référence pour les deux conditions de débit dans la bande de fréquence thêta (4-7 Hz) qui englobe les fluctuations lentes d'amplitude de l'enveloppe temporelle de la parole. Les moyennes de chaque groupe (DYS et NT) dans chaque condition de débit ont été calculées, et des statistiques non paramétriques avec la méthode de Monte-Carlo (1000 répétitions, correction au niveau des clusters) ont été réalisées sur 4 ROIs bilatérales définies selon les modèles de la parole (Giraud, Poeppel, 2012; Hickok, Poeppel, 2007) : gyrus de Heschl, gyrus temporal supérieur, gyrus temporal médian et cortex précentral (atlas *Automated Anatomical Labeling* ou AAL). Nous avons également conduit des analyses de

corrélation de Spearman entre la cohérence cortico-acoustique au sein des ROIs et les performances langagières des enfants DYS et NT telles que mesurées par la BALE et l'ELO (cf. partie 2.1).

# 3 Résultats

La FIGURE 1 présente les cartes de cohérence cortico-acoustique dans la bande thêta (4-7 Hz) chez les enfants dysphasiques (1A) et les enfants neurotypiques (1B) dans les conditions de débit normal et de débit rapide, ainsi que le contraste direct entre les deux groupes (1C). Lors de la perception de parole à débit normal, chez les enfants NT, une augmentation de cohérence entre le signal cortical MEG et l'enveloppe d'amplitude des phrases est principalement observée dans le cortex temporal médian et supérieur antérieur droit ainsi que dans le cortex fronto-pariétal droit. Chez les enfants DYS en revanche, ce pattern n'est pas retrouvé. Les analyses statistiques montrent en effet un couplage cortico-acoustique réduit chez ces enfants, comparés à leurs pairs au développement typique, au sein des gyri temporaux médian et supérieur antérieurs droits (Aires de Brodmann BA 21 et 22 ; 1D, panneau de gauche) en condition de débit normal. Notons que ceci est observé de manière bilatérale lorsque la correction statistique n'est pas appliquée (analyses non corrigées, 1E, panneau de gauche). Lors de la perception de parole rapide, aucune différence significative entre les groupes n'est observée pour les analyses corrigées. Cependant, les résultats non corrigés (1E, panneau de droite) révèlent un alignement plus fort des oscillations thêta sur le signal acoustique dans le cortex prémoteur gauche (BA 6 et 8), la partie antérieure du cortex temporal supérieur droit (BA 22) et le cortex auditif primaire droit (BA 41) chez les enfants NT comparés aux enfants DYS. Les enfants DYS montrent quant à eux une augmentation de cohérence au niveau du cortex temporal supérieur postérieur droit (BA 22) et des gyri angulaire et supramarginal droits (BA 39 et 40) dans cette condition de débit, alors qu'elle n'est pas observée chez les enfants NT.

Il faut par ailleurs noter que les analyses des puissances dans la bande thêta n'ont montré aucune modulation significative (corrigée ou non) de la puissance dans les régions corticales précitées lors de la perception de parole à débit normal ou rapide (par rapport à la période de référence), chez aucun des deux groupes d'enfants. Ceci suggère que les augmentations de cohérence cortico-acoustique observées reflètent bien un processus de synchronisation oscillatoire et non simplement une augmentation de l'activité corticale.

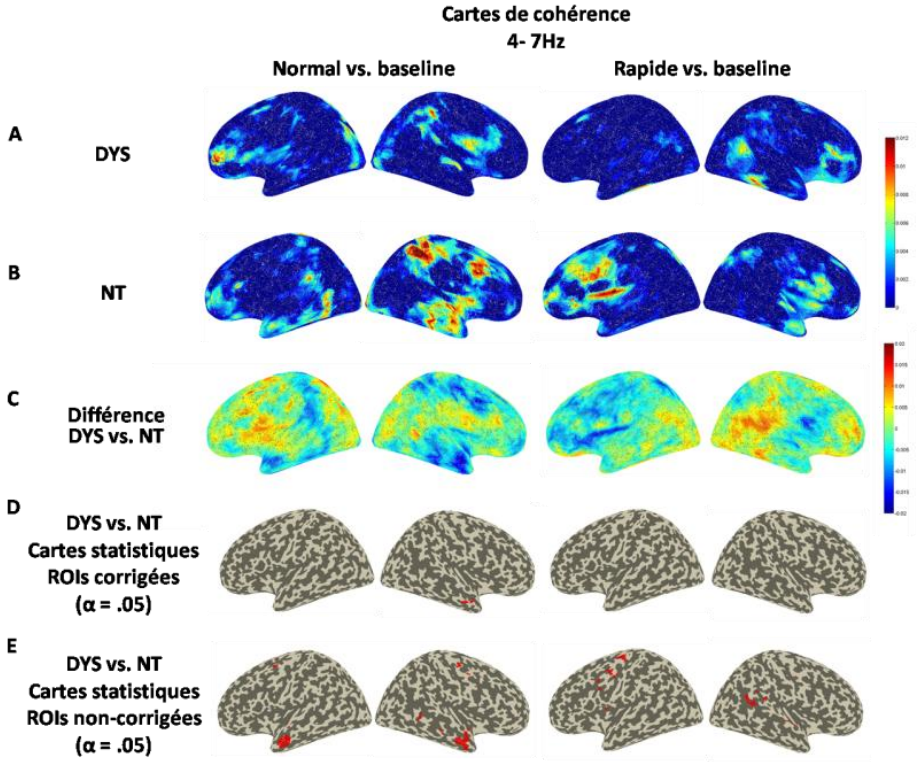



FIGURE 1 : Cartes de cohérence cortico-acoustique dans la bande thêta (4-7 Hz) lors de la perception de parole à débit normal (panneau de gauche) et à débit rapide (panneau de droite), par rapport à la ligne de base (période de référence ou *baseline*), chez les enfants dysphasiques (A) et les enfants neurotypiques (B). Le contraste direct entre les deux groupes est présenté en C. Les cartes statistiques corrigées (D) et non corrigées (E) de cette différence sont également reportées.

Les analyses de corrélation entre la cohérence cortico-acoustique dans la bande thêta (4-7 Hz) et les performances langagières des enfants ne montrent aucune corrélation significative chez les enfants NT. En revanche, chez les enfants DYS, elles révèlent d'une part une corrélation négative entre la cohérence dans la région temporale médiane postérieure droite (incluant le gyrus angulaire) et les performances phonologiques des enfants (tâches de répétition et métaphonologiques ; $r = -0.75$, $p < .01$, FIGURE 2A). D'autre part, une corrélation positive est observée entre la cohérence dans le cortex temporal supérieur postérieur droit (englobant le gyrus supramarginal) et le propre débit de parole des enfants mesuré dans une tâche narrative ($r = 0.91$, $p < .001$, FIGURE 2B). Autrement dit, plus les enfants DYS sont capables de parler vite, plus le cortex temporal postérieur droit synchronise son activité oscillatoire sur l'enveloppe d'amplitude des signaux de parole rapide. Ce couplage cortico-acoustique est par ailleurs d'autant plus présent que les enfants ont des performances phonologiques faibles. Dans l'ensemble, ces corrélations suggèrent donc que l'augmentation de cohérence dans le cortex temporal postérieur droit chez les enfants dysphasiques reflète un pattern atypique et la mise en place possible d'un phénomène compensatoire pour traiter la parole rapide.

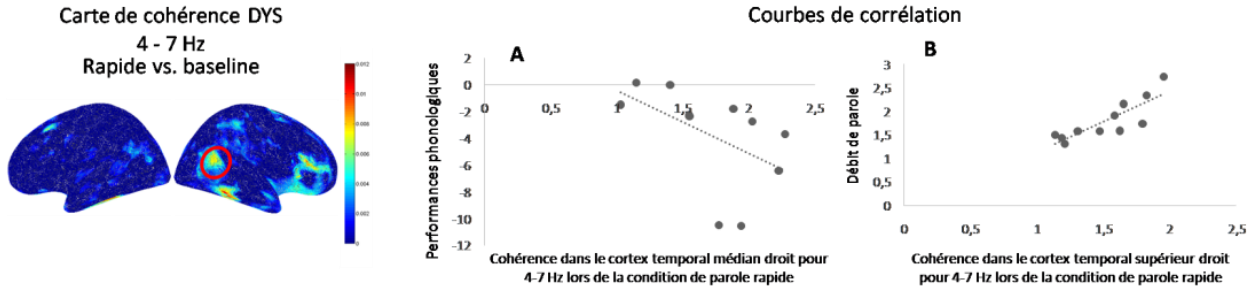



FIGURE 2 : Corrélations de Spearman (rho) entre (A) la valeur moyenne de cohérence cortico-acoustique dans la bande thêta (4-7 Hz) dans le cortex temporal médian postérieur droit lors de la perception de parole rapide et les performances phonologiques des enfants DYS, et (B) la valeur moyenne de cohérence en 4-7 Hz dans le cortex temporal supérieur droit dans la condition rapide et le propre débit de parole des enfants. Les valeurs sont exprimées en scores Z.

# 4 Discussion

Les résultats dans la condition de parole à débit normal suggèrent un alignement plus faible des oscillations corticales thêta dans le cortex auditif associatif droit sur l'enveloppe d'amplitude du signal chez les enfants dysphasiques, comparés à des enfants neurotypiques du même âge. Nos observations chez les enfants neurotypiques s'accordent avec les modèles oscillatoires ainsi qu'avec les études chez l'adulte (Giraud, Poeppel, 2012) montrant un couplage cortico-acoustique dans la bande thêta préférentiellement dans le cortex auditif droit, qui serait plus à même de segmenter la parole en unités syllabiques. Les rares études menées chez l'enfant, utilisant des stimuli non verbaux ou verbaux très simples, ont également décrit ce pattern de latéralisation (Abrams et al., 2008). Notre étude révèle qu'alors que les enfants neurotypiques âgés de 8 à 13 ans synchronisent leurs oscillations corticales sur le rythme syllabique de phrases naturellement produites à un débit normal, ceci ne semble pas être le cas chez les enfants dysphasiques. Les régions auditives droites des dysphasiques seraient ainsi moins capables de suivre les fluctuations lentes de l'enveloppe d'amplitude de la parole, même produite à un débit normal, ce qui conduirait à une extraction moins efficace de l'information rythmique (syllabique ici) et par conséquent au développement de représentations phonologiques et morphosyntaxiques imprécises. Dans l'ensemble, nos données s'accordent donc avec les prédictions de la « *prosodic phrasing hypothesis* » (Cumming et al., 2015) et les études montrant l'existence de déficits de traitement du rythme chez les dysphasiques (Corriveau et al., 2007).

Lorsque la parole est accélérée naturellement, des réseaux corticaux distincts et présentant une latéralisation hémisphérique différente semblent être engagés chez les deux groupes d'enfants (rappelons que les résultats étaient significatifs sans correction statistique et mériteraient donc d'être confirmés avec un plus grand échantillon). Alors que les régions (pré)motrices gauches se synchronisent sur la parole rapide chez les enfants neurotypiques, le couplage cortico-acoustique est observé dans le cortex temporal postérieur droit chez les enfants dysphasiques. Les travaux chez l'adulte montrent une implication spécifique des régions motrices gauches ainsi qu'une connectivité fonctionnelle renforcée entre les régions auditives et motrices lors de la perception de parole dans des conditions difficiles (Alho et al., 2014). Nos données chez l'enfant neurotypique révèlent que

lorsque la parole est produite plus rapidement (i.e. plus difficile à décoder), l'activité oscillatoire des régions de production de la parole se synchronise sur les modulations d'enveloppe du signal. Ceci est en accord avec l'existence d'une voie dorsale d'intégration sensori-motrice (modèle à double voie ; Hickok, Poeppel, 2007) ainsi qu'avec les études suggérant un rôle prédictif des régions articulatoires dans le traitement de l'information auditive temporelle (Morillon et al., 2014). Chez les enfants dysphasiques, nous n'avons observé aucun couplage cortico-acoustique dans le cortex (pré)moteur gauche, évoquant un dysfonctionnement possible de la voie dorsale sensori-motrice lorsqu'il s'agit de s'aligner sur un rythme syllabique rapide. Ceci pourrait rendre compte du déficit de perception de parole rapide que nous avions précédemment mis en évidence chez un autre groupe d'enfants dysphasiques (Guiraud et al., 2018). Nos résultats s'accordent également avec les travaux montrant que les troubles rythmiques dans la dysphasie s'expriment non seulement au niveau perceptif mais aussi moteur (i.e. tâches de *tapping*) et qu'ils pourraient donc refléter un dysfonctionnement du couplage auditivo-moteur (Corriveau, Goswami, 2009). Par ailleurs, chez les dysphasiques, nous avons mis en évidence une synchronisation des oscillations du cortex temporal postérieur droit, s'étendant à la jonction temporo-pariétale, sur la parole rapide, qui n'était pas retrouvée chez les enfants neurotypiques. Cette région appartient à la voie dorsale sensori-motrice, normalement latéralisée à gauche, et est censée participer aux processus phonologiques à la fois en perception et en production. Il est notamment intéressant de noter que des études neuroanatomiques et fonctionnelles ont révélé une absence de latéralisation à gauche (observée dans le développement typique), voire une latéralisation à droite, de cette région chez les dysphasiques (Badcock et al., 2012). Nos résultats, s'ils sont confirmés avec un plus grand échantillon, pourraient alors suggérer que l'implication de la région temporale postérieure droite lors de la perception de parole rapide chez les enfants dysphasiques reflète un phénomène compensatoire. La voie dorsale sensori-motrice pourrait être fonctionnelle mais présenterait une latéralisation hémisphérique différente de celle observée chez les enfants neurotypiques (voir Cutini et al., 2016 chez des enfants dyslexiques). Cette interprétation semble corroborée par les résultats préliminaires de nos analyses de corrélation : si l'activité du cortex temporal postérieur droit reflète des mécanismes d'intégration sensori-motrice, la plus forte synchronisation oscillatoire dans cette région sur la parole rapide chez les enfants dysphasiques capables de parler vite pourrait indiquer que le couplage auditivo-moteur (même latéralisé à droite) se produise dans une certaine mesure. Autrement dit, la région temporale droite d'intégration sensori-motrice serait plus à même de décoder la parole rapide chez les enfants dysphasiques présentant moins de troubles expressifs. Néanmoins, l'implication de cette région droite était particulièrement observée chez les enfants présentant des troubles phonologiques sévères (corrélation négative). Bien qu'aucune relation causale ne puisse être inférée à partir de nos analyses actuelles, ceci suggère que nos résultats reflètent un pattern atypique, en termes d'asymétrie hémisphérique notamment, du traitement de la parole dans la dysphasie.

En conclusion, notre étude apporte, pour la première fois à notre connaissance, des éléments en faveur d'un alignement atypique entre rythme syllabique et rythmes cérébraux lors de l'écoute de parole naturelle à débit normal et rapide chez les enfants dysphasiques.

# Remerciements

ANR ODYSSEE (PI : V.B., n°11 JSH2 005 1) ; LabEx ASLAN (ANR-10-LABX-0081, bourse de doctorat : H.G.) et *Erasmus Mundus Auditory Cognitive Neuroscience* (H.G.) ; Programme des Chaires du Canada et de la Subvention à la découverte du CRSNG (K.J. ; RGPIN-2015-04854).

# Références


ABRAMS D., NICOL T., ZECKER S., KRAUS N. (2008). Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. *Journal of Neuroscience* 28(15), 3958-3965.

ALHO J., LIN F.-H., SATO M., TIITINEN H., SAMS M., JÂÂSKELÄINEN I. P. (2014). Enhanced neural synchrony between left auditory and premotor cortex is associated with successful phonetic categorization. *Frontiers in Psychology* 5, 394.

BADCOCK N. A., BISHOP D. V. M., HARDIMAN M. J., BARRY J. G., WATKINS K. E. (2012). Co-localisation of abnormal brain structure and function in specific language impairment. *Brain and Language* 120(3), 310-320.

CORRIVEAU K. H., PASQUINI E. S., GOSWAMI U. (2007). Basic auditory processing skills and specific language impairment: a new look at an old hypothesis. *Journal of Speech Language and Hearing Research* 50(3), 647-666.

CUTINI, S., SZUCS, D., MEAD, N., HUSS, M., GOSWAMI, U. (2016). Atypical right hemisphere response to slow temporal modulations in children with developmental dyslexia. *NeuroImage*, 143, 40-49.

CORRIVEAU K. H., GOSWAMI U. (2009). Rhythmic motor entrainment in children with speech and language impairments: tapping to the beat. *Cortex* 45, 119-130.

CUMMING R., WiILSON A., GOSWAMI U. (2015). Basic auditory processing and sensitivity to prosodic structure in children with specific language impairments: a new look at a perceptual hypothesis. *Frontiers in Psychology* 6, 972.

GHITZA O. (2011). Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology* 2, 130.

GHITZA O., GREENBERG S. (2009). On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica* 66(1–2), 113-126.

GIRAUD A.-L., POEPPEL, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience* 15(4), 511-517.

GUIRAUD H., BEDOIN N., KRIFI-PAPOZ S., HERBILLON V., CAILLOT-BASCOUL A., GONZALEZ-MONGE S., BOULENGER V. (sous presse). Don't speak too fast! Processing of fast rate speech in children with Specific Language Impairment. *PLoS ONE*.

HICKOK G., POEPPEL D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience* 8(5), 393-402.

MAILLART C., SCHELSTRAETE M. A. (2012). *Les dysphasies: de l'évaluation à la rééducation.* Issy-les-Moulineaux, France : Elsevier-Masson.

MOLINARO N., LIZARAZU M., LALLIER M., BOURGUIGNON M., CARREIRAS M. (2016). Out-of-synchrony speech entrainment in developmental dyslexia. *Human Brain Mapping* 37(8), 2767-2783.

MORILLON B., SCHROEDER C. E., WYART V. (2014). Motor contributions to the temporal precision of auditory attention. *Nature Communications 5,* 1-9.

PEELLE J. E., GROSS J., DAVIS, M. H. (2013). Phase-Locked Responses to Speech in Human Auditory Cortex are Enhanced During Comprehension. *Cerebral Cortex* 23(6), 1378-1387.

# Perception et production de /y/ et /u/ en français L2 chez l'apprenant anglophone débutant : étude de cas de leur catégorisation chez quatre locuteurs

Delfine Michaud[1], Nicolas Ballier[1]
(1) CLILLAC-ARP (EA3967), 5 rue Thomas Mann, 75205 Paris Cedex 13
delfine.michaud@yahoo.fr, nicolas.ballier@univ-paris-diderot.fr

## Résumé

Cette étude aborde la prononciation de la voyelle antérieure haute et arrondie /y/ du français par des locuteurs d'anglais américain apprenant le français en L2. Le contraste de labialisation n'existant pas en anglais, cette étude cherche à montrer comment /y/ est catégorisé dans les débuts de l'interlangue des locuteurs américains, c'est-à-dire si cette voyelle est considérée comme « similaire » ou « nouvelle » par les apprenants L2 (Major 2001, Flege 2002, Birdsong 2003, Colantoni, Steele, et Escudero 2015). Cet article expose les résultats de l'expérience menée, pour laquelle quatre étudiants américains ont été enregistrés au début de leur apprentissage du français et ont participé à un test de discrimination. Quatre francophones natifs ont ensuite participé à un test d'assimilation afin de juger leurs productions. Une analyse de l'influence de différents facteurs sur la production des voyelles françaises hautes et arrondies est également proposée.

## Abstract

**Categorizing lip-rounding in /y/ and /u/ for four American early learners of French**
This study aims at addressing the issues that American-English speaking learners of L2 French are faced with when attempting to produce French high front rounded vowel /y/. As the contrast of rounding exists in French but not in English, this study looks at the way /y/ is categorised in the interlanguage of 4 AmE speakers, namely whether it is thought of as a similar or as a dissimilar sound by L2 learners (Major 2001, Flege 2002, Birdsong 2003, Colantoni, Steele, & Escudero 2015). It discusses the results of the experiment that was carried out, for which 4 American university students were recorded in an early stage of their L2 French learning and participated in a perceptual discrimination test. 4 French native speakers then participated in a perceptual assimilation test in order to judge the AE speakers' productions. The influencing factors on the production of French high rounded vowels are also analysed.

Mots-clés : Interphonologie, labialité, FLE, perception, production
Keywords: Interphonology, lip-rounding, L2 French, perception, production.

# 1 Introduction et contexte

Cette étude s'intéresse au comportement acoustique des voyelles hautes du français produites par des apprenants débutants du français L2 de nationalité américaine. Alors qu'en anglais américain il n'existe pas de voyelles antérieures arrondies, le français compte la labialisation parmi les traits distinctifs de ses voyelles antérieures (Levy 2009). Nous pouvons donc nous attendre, de la part d'apprenants de français L2 anglophones, à une production problématique des voyelles antérieures arrondies du français /y/, /ø/ et /œ/. Compte tenu de ce que le contraste fondé sur la labialité des voyelles d'avant n'existe pas en anglais, la voyelle française /y/ peut être considérée comme une


10.21437/JEP.2018-27

voyelle « nouvelle » pour les apprenants anglophones du français L2. Au contraire, la voyelle /u/ apparaît dans l'inventaire des voyelles des deux langues – malgré des propriétés acoustiques différentes – et elle peut ainsi être considérée comme une voyelle « similaire » pour les apprenants anglophones du français L2.

Cette interprétation est en revanche démentie par certaines études, telle que celle de (Mayr et Escudero, 2010), qui explique que des apprenants anglophones de l'allemand L2, langue qui comprend la voyelle antérieure arrondie /y/ - quoiqu'acoustiquement différente du /y/ français - assimilent cette voyelle à un /u:/ anglais dans 98% des cas, et assimilent le /u/ allemand à une autre voyelle que le /u:/ anglais dans 20% des cas (Mayr et Escudero 2010). Une autre étude, cette fois proposée par Flege, établit que le /y/ français est acoustiquement plus proche du /u:/ anglais que le /u/ français ne l'est du même son (Colantoni et al. 2015). Leur conclusion est donc que /y/ ne serait pas une voyelle « nouvelle » mais bien une voyelle « similaire ».

Si l'on en croit le *Speech Learning Model* (SLM), modèle développé par Flege en 1995, un son « nouveau » s'acquiert plus facilement qu'un son « similaire », car plus un son est éloigné, ou « dissemblable », d'un autre son, plus il est facile pour un apprenant L2 de créer une nouvelle catégorie phonologique (Major 2001, Flege 2002, Birdsong 2003). Par conséquent, si, au sein de la paire /y/-/u/, /y/ est la voyelle « nouvelle » et /u/ est la voyelle « similaire », alors /y/ devrait être plus facile à acquérir que /u/, et vice versa (Flege 2002, Escudero 2007).

Notre étude cherche en ce sens à tester le modèle du SLM sur ce point et cherche à répondre à plusieurs questions. D'une part, qu'en est-il de la catégorisation de /y/ et /u/ par des apprenants débutants de français L2 anglophones ? D'autre part, nous avons voulu évaluer les facteurs possibles d'influence sur l'exactitude de la production des apprenants L2, notamment la tâche demandée aux apprenants et le temps d'exposition et de pratique de la L2. Nous détaillerons ici la méthodologie ainsi que les résultats, que nous discuterons ensuite.

# 2 Méthodologie

L'expérience s'est déroulée en trois temps, dont les deux premiers étaient identiques l'un à l'autre mais espacés de six mois. Dans les deux premiers temps (phases 1 et 2), nous avons testé la perception des locuteurs d'anglais américain de paires de voyelles françaises grâce à un test de discrimination, puis nous avons enregistré leur production de mots et de phrases isolées. Le troisième temps a consisté à proposer un second test de perception, cette fois basé sur l'assimilation des voyelles produites par les locuteurs américains, et soumis à des locuteurs natifs du français.

## 2.1 Sujets et corpus

Pour l'expérience, nous avons choisi quatre locuteurs natifs de l'anglais américain (deux femmes et deux hommes) et quatre locuteurs natifs du français (deux femmes et deux hommes). Les locuteurs de l'anglais américain étaient tous étudiants à l'Université du Minnesota, aux États-Unis, et ont suivi des cours de français durant la totalité de l'expérience, à raison de quatre heures présentielles par semaine avec une pause aux vacances d'hiver. La première partie de l'expérience s'est déroulée moins de trois mois après le début des cours, et la seconde partie six mois plus tard. Une seule étudiante avait déjà suivi des cours de français L2 pendant un an au lycée, plusieurs années auparavant. Tous avaient déjà suivi des cours d'une autre langue étrangère par le passé. Les locuteurs du français, de leur côté, jugeaient leur utilisation de la langue française entre 90 et 100% de leur temps d'utilisation des langues au moment de l'expérience, avaient tous déjà étudié au moins une langue étrangère (estimaient par ailleurs que leur niveau d'anglais était au minimum B1), et avaient tous une formation universitaire.

Les apprenants anglophones devaient, pour chacune des deux premières phases de l'expérience, discriminer des paires minimales du français (perception), lire une liste de mots et de phrases en français (production) et répéter une liste de mots, phrases et logatomes en français (production). Les mots et phrases utilisés pour tester la perception et la production des apprenants recouraient principalement aux sons /i/, /y/ et /u/ du français, en contextes ouverts et fermés. Ces listes de mots, phrases et logatomes ont été créées spécialement pour l'expérience et étaient inspirés des travaux de Akyüz, Bazelle-Shahmaei, Bonenfant, Flament, Lacroix, Moriot et Renaudineau (2002) et de Abry et Chalaron (2011).

Le groupe de natifs francophones (experts) devait, lors de la troisième phase, classer les sons isolés des productions des apprenants anglophones selon les sons du français, phase utile pour juger la catégorisation des sons par les apprenants. Un total de 169 stimuli, sélectionnés à partir de 44 mots produits par les apprenants L2, a été utilisé pour le test d'assimilation proposé au groupe de locuteurs natifs français. Tous portaient sur les voyelles /i/, /y/ et /u/.

### 2.2 Procédure

Les deux premières phases de l'expérience ont eu lieu dans un studio d'enregistrement au sein d'une université américaine, et les production des apprenants anglophones ont été enregistrées à l'aide d'un Marantz Professional PMD 620 MKII, avec une fréquence d'échantillonnage de 44,1 kHz et une résolution de 16 bits. Les listes servant de support de lecture et de répétition étaient affichées sur un écran d'ordinateur. Les différents tests de perception ont été réalisés à l'aide du logiciel *Praat* (Boersma et Weenink 2017) à l'aide d'un script créant une expérience de type *Multiple Forced Choice* (MFC) durant laquelle les stimuli étaient proposés dans un ordre aléatoire. Après écoute, les participants devaient décider :
- pour le test de discrimination ABX, si les sons qu'ils entendaient étaient « similaires » ou « différents »,
- pour le test d'assimilation, si le son qu'ils entendaient était /i/, /y/, /u/, /e/, /ø/, /o/ ou /œ/ (les symboles phonétiques n'étaient pas utilisés). Pour chaque type de test, les participants pouvaient réécouter les stimuli une seule fois et n'obtenaient aucun feedback.

### 2.3 Extraction des résultats

L'extraction des réponses aux tests de perception a été faite sous *Praat*. Les fichiers audios des productions L2 ont eux aussi été annotés avec *Praat* et découpés avec *Audacity* (Audacity 2008). Des scripts *Praat* créés par Cédric Gendrot ont ensuite été utilisés pour, d'une part, extraire l'information acoustique des voyelles, et d'autre part, pour dessiner des triangles vocaliques (Gendrot s.d.). L'ensemble des résultats a enfin été analysé sous *R* (R Core Team 2017).

## 3 Résultats

### 3.1 Perception : peu de difficulté

Les résultats des tests de perception des apprenants L2 montrent peu de difficultés concernant la discrimination des sons /y/ et /u/, avec une moyenne de réussite de 91,7% au début de leur apprentissage, et de 97,9% six mois plus tard. De même, lorsque les mêmes apprenants étaient soumis à deux stimuli identiques de /y/, le taux de réussite était de 95% en début d'expérience et de 97,5% lors de la deuxième phase. La discrimination des sons /y/ et /i/ a montré encore moins de difficultés, avec respectivement 98,4% et 100% de perception correcte pour la première phase et la seconde phase de l'expérience.

### 3.2 Production : /y/ et /u/ problématiques

L'analyse acoustique montre que les valeurs F2 du **/u/** français produit par les anglophones indiquent une tendance à produire ce phone plus en avant que le /u/ en français L1 (beaucoup de valeurs F2 sont comprises entre 1000 et 1700 Hz, contre une moyenne de 850 Hz en français L1 (Gendrot et Adda-Decker 2005)). Pour **/y/**, les valeurs de F2 oscillent sur une large échelle (1000 à 2400 Hz), entre celles du /y/ français et celles du /u:/ anglais. De même, les valeurs de F3 pour /y/ se trouvent entre celles du /y/ français et du /u:/ anglais.

Pour identifier le rôle respectif des variables dans la prédiction de l'identification des voyelles par les experts, nous avons procédé à une régression logistique à effets mixtes à l'aide du package {lme4} (Bates, Mächler, Bolker, & Walker, 2015). Nous avons considéré que la classe de voyelle à prononcer, la tâche demandée, le contexte consonantique et la phase de l'expérience étaient les effets fixes et que les mots étaient l'effet aléatoire. En résumé, les résultats de ce modèle mixte permettent de mettre en évidence que les voyelles /u/ et /y/ avaient respectivement 22,99 et 74,22 fois moins de chances d'être correctement perçues par les experts que la voyelle /i/ ($p < 0{,}001$). Le tableau ci-dessous (Table 1) montre en efet un très haut taux de perception correcte pour /i/ (94,35%), malgré quelques classements en /e/ probablement dues à des F1 un peu trop hauts. La voyelle /u/ a été correctement perçue dans 56,97% des cas, mais les résultats montrent aussi un nombre important d'assimilations en /ø/ (23,77%), ainsi qu'un moindre pourcentage d'assimilations en /y/ (8,20%). Des valeurs F1 et F2 plus hautes que des valeurs natives pourraient expliquer ces assimilations. Enfin, la voyelle /y/ a été correctement perçue dans 34,09% des cas, ce qui ne représente pas une majorité. En effet, /y/ a été perçu comme un /u/ dans 36,04% des cas, ce que l'on peut facilement expliquer par des valeurs F2 plus hautes qu'en français, correspondant au /u:/ anglais. Les assimilations en /ø/ sont fréquentes elles aussi (23.38%) ; elles correspondent à des valeurs F2 plus basses qu'en français, et à quelques valeurs F1 un peu hautes, que l'on peut observer sur les résultats de l'analyse acoustique des productions L2.

| | e | ø | i | o | œ | u | y | Total |
|---|---|---|---|---|---|---|---|---|
| i | 4.84 | 0.00 | **94.35** | 0.00 | 0.00 | 0.00 | 0.81 | 100 % |
| u | 0.82 | 23.77 | 0.00 | 6.97 | 3.28 | **56.97** | 8.20 | 100 % |
| y | 0.00 | 23.38 | 0.32 | 2.60 | 3.57 | 36.04 | **34.09** | 100 % |

TABLE 1 : Matrice de confusion (en pourcentage d'assimilation) des voyelles perçues par les experts.

### 3.3 Influences de la pratique de la L2 et de la tâche demandée

De manière générale, la pratique de la L2, traduite ici par l'intervalle de six mois entre les phases 1 et 2 de l'expérience, ainsi que la tâche demandée aux apprenants (lecture ou répétition des segments) a eu un impact sur la façon dont le groupe de francophones natifs assimilait les sons entendus. Le taux de réussite du test de perception du groupe d'experts passe de 47,29% à 59,39% de la première phase à la deuxième phase, et il est de 39,33% en lecture, contre 69,06% en répétition.

Tout d'abord, le temps passé à l'étude de la L2. Si l'on y regarde de plus près, il est 2,24 fois plus probable que le groupe d'experts perçoive correctement les voyelles produites lors de la deuxième phase que de la première ($p < 0{,}001$) ; ceci est plus particulièrement visible si l'on ne se concentre que sur les tâches de lecture, durant laquelle les apprenants L2 n'étaient pas influencés : il existe en

effet une différence significative dans la perception des voyelles qui ont uniquement été produites en tâche de lecture ($p < 0.001$). En revanche, l'observation de la distribution des valeurs F1 et F2 suite à l'extraction des résultats n'a pas montré de différence significative entre les deux phases, une absence de différence particulièrement manifeste en tâche de lecture (Figure 2). Seule l'observation des valeurs F3 de /y/ indique une évolution vers plus d'authenticité dans la labialisation de la voyelle d'une phase d'enregistrement à l'autre (Figure 3).

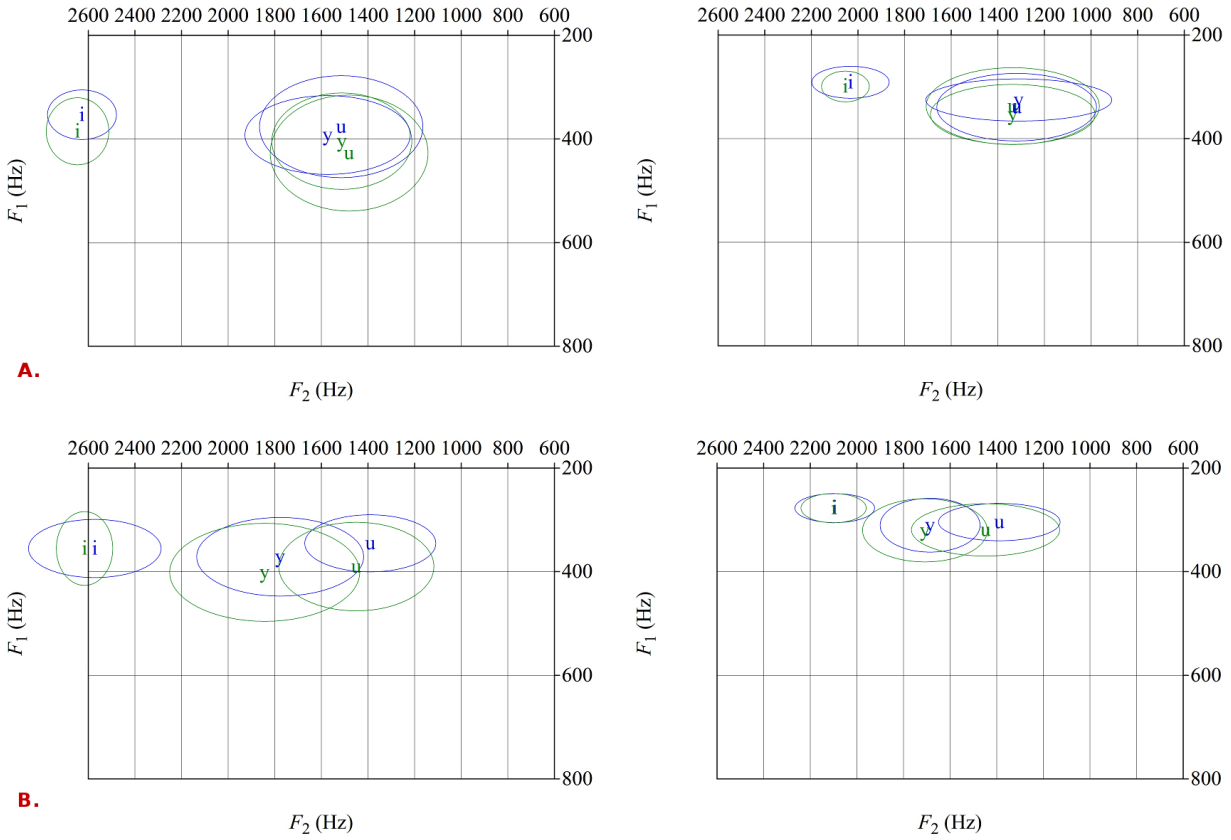



FIGURE 2: Distribution de /i/, /y/ et /u/ sur des triangles vocaliques représentant F1 et F2. Valeurs de la première phase en vert, valeurs de la deuxième phase en bleu. (Femmes à gauche, hommes à droite. A : lecture. B : répétition)

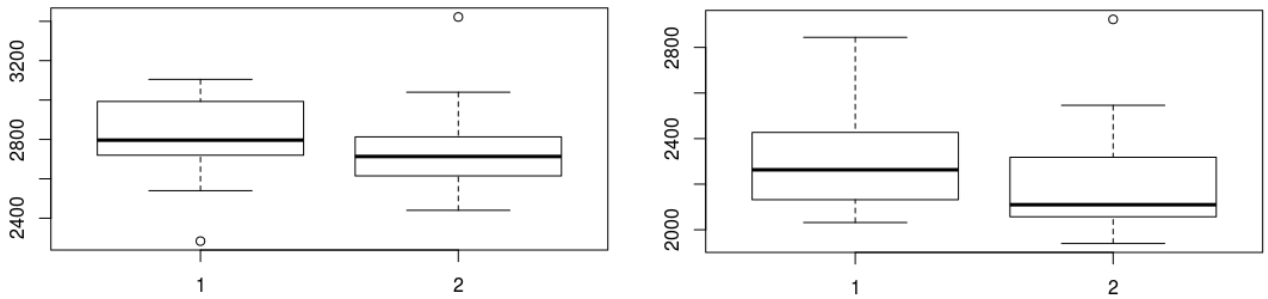



FIGURE 3: Boites à moustaches représentant la distribution des valeurs F3 de /y/ en phases 1 et 2. (Femmes à gauche, hommes à droite).

S'agissant de la tâche demandée aux apprenants L2, le test d'identification de la voyelle prononcée auquel ont été soumis les experts a montré qu'il y avait 5,26 fois plus de chances pour que la perception des voyelles soit correcte en tâche de répétition plutôt qu'en tâche de lecture ($p < 0{,}001$). La différence de réussite dans la perception est d'autant plus significative pour les voyelles /y/ et /u/, qui sont mieux perçues en répétition qu'en lecture ($p < 0{,}001$). Les données acoustiques collectées dans les productions L2, qui sont ici des valeurs brutes sans normalisation, montrent également une hausse des valeurs F2 pour /y/ en répétition, démarquant cette voyelle de /u/ en lui donnant sa caractéristique antérieure (voir les triangles vocaliques de la Figure 4).

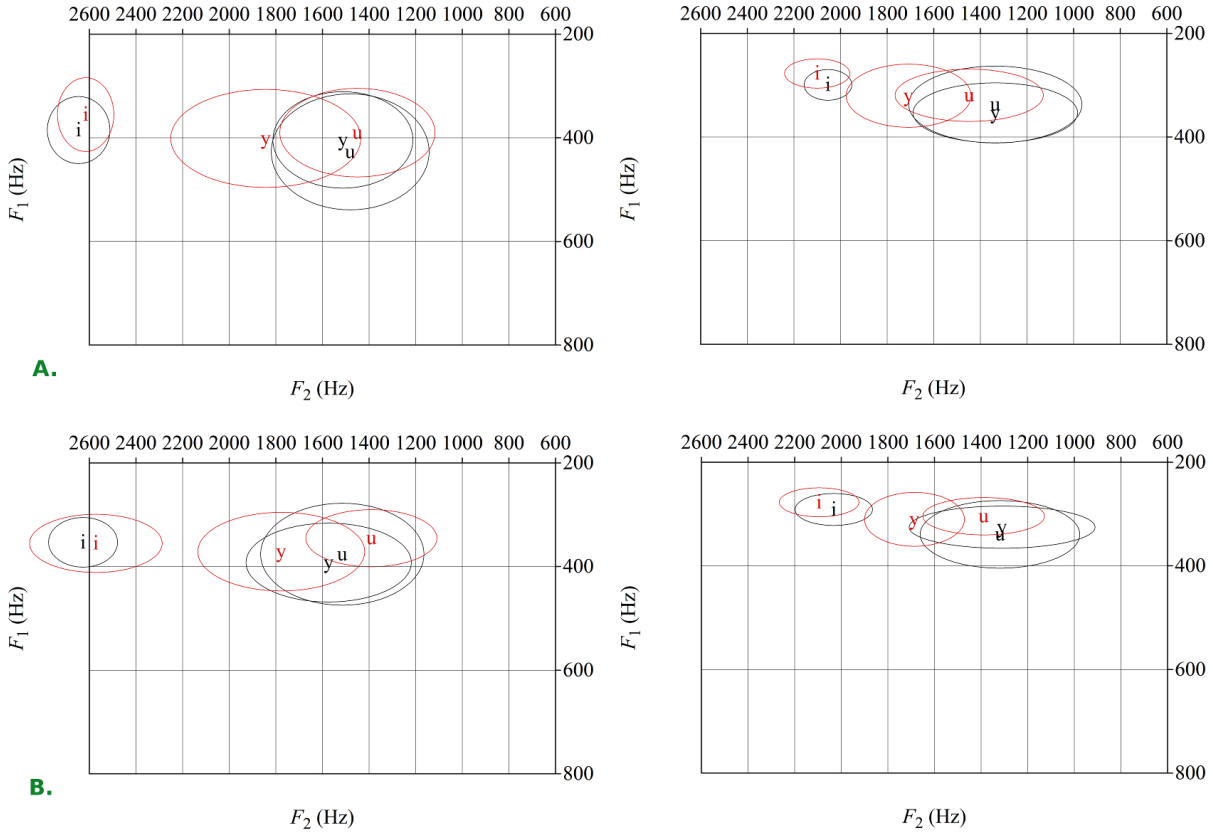



FIGURE 4: Distributions de /i/, /y/ et /u/ sur des triangles vocaliques représentant F1 et F2. Valeurs de la tâche de lecture en noir ; valeurs de la tâche de répétition en rouge. Femmes à gauche, hommes à droite. A : phase 1. B : phase 2.

# 4 Discussion et conclusion

Les résultats des tests de discrimination montrent une grande majorité de réponses correctes, ce qui rejoint les conclusions de (Best, Faber, et Levitt 1996) sur le fait que les apprenants débutants dont la langue maternelle est l'anglais américain ne rencontrent pas de grandes difficultés lors de la discrimination de paires comprenant /y/ et une autre voyelle. Si l'on compare les résultats de la discrimination de /y/-/u/ et de /y/-/i/, l'on remarque que /y/ est plus facilement confondu avec /u/ qu'avec /i/, ce qui montre que la voyelle /y/ est plus facilement perçue par les apprenants L2 comme une voyelle arrondie que comme une voyelle antérieure.

L'analyse des données acoustiques des apprenants indique que les valeurs F2 de /y/ ont tendance à ressembler à celles de /u:/ en anglais, ce qui a mené le groupe de francophones natifs à ne pas reconnaître /y/ dans une grande partie des cas. Ces résultats corroborent l'étude déjà mentionnée de (Mayr et Escudero 2010), qui observait l'assimilation du /y/ allemand au /u:/ anglais 98% du temps. Notre étude montre que nos apprenants semblent avoir catégorisé /y/ comme la voyelle /u:/ anglaise et non comme une nouvelle voyelle, et qu'ils n'ont pas catégorisé /u/ comme la voyelle /u:/ anglaise (ce qui est confirmé par le test de discrimination).

Enfin, le modèle mixte a montré que la production des apprenants L2 avait été influencée par des facteurs tels que le temps de pratique de la L2 et la tâche demandée. La perception du groupe d'experts montre que le temps de pratique entre les deux phases a eu un impact positif sur l'authenticité de la prononcitation des apprenants L2. Le rôle de la tâche demandée aux apprenants est considérable, puisqu'en répétition les valeurs F2 de /y/ se rapprochent des valeurs natives, et que la perception des experts en est significativement meilleure. Pour déterminer plus précisément le rôle de chaque facteur dans l'identification des voyelles, nous avons eu recours à un arbre d'inférence conditionnelle, réalisé à l'aide du package de R {party}. La Figure 5[1] permet d'observer l'influence, plus ou moins significative, de chaque facteur (voyelle, tâche et phase de l'expérience[2]) sur le taux de réussite du groupe d'experts lors de l'assimilation des voyelles. On peut clairement voir sur le nœud 1 qu'il y a une différence significative ($p < 0{,}001$) en fonction de la voyelle prononcée par les apprenants (/i/ d'un côté, /y/ et /u/ de l'autre). Le nœud 2 montre une différence significative ($p < 0{,}001$) en fonction de la tâche réalisée (lecture d'un côté, et répétition de l'autre). Les taux de réussite s'affichent dans les boîtes en bas de l'arbre.

Pour conclure, cette étude semble valider certaines des hypothèses du SLM. Un son de la L2 considéré comme « similaire » à un son de la L1 paraît plus difficile à acquérir, comme c'est le cas

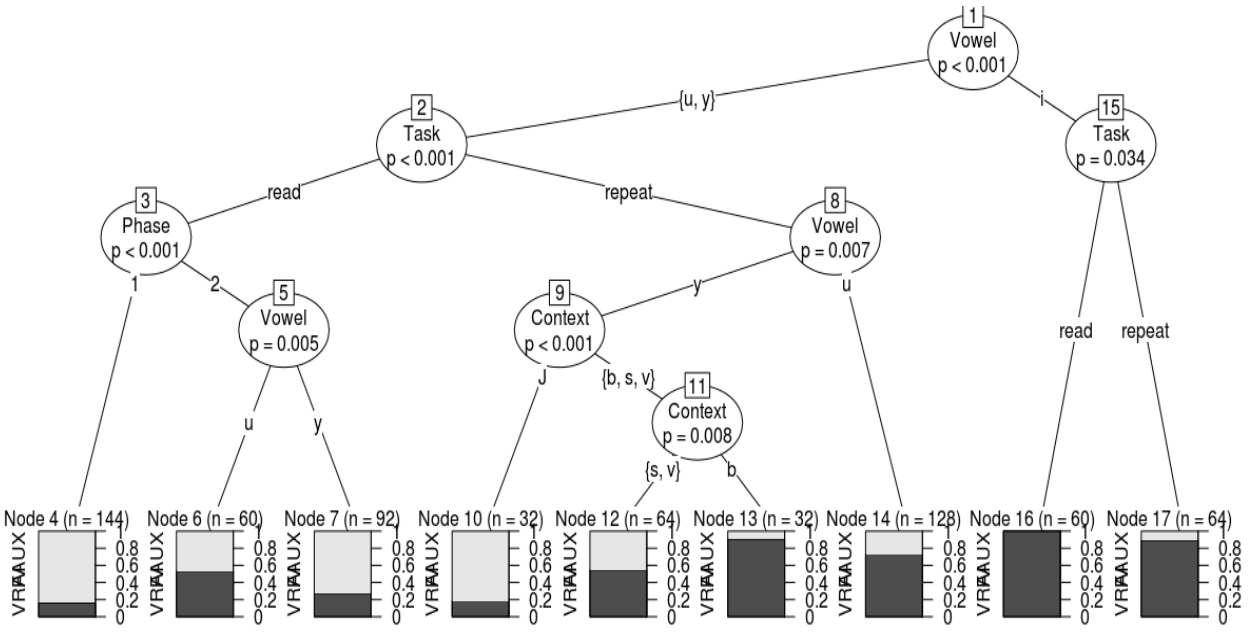



FIGURE 5: CTree (arbre d'inférence conditionnelle) montrant le taux de réussite au test de perception des experts. Pour chaque nœud, apparaissent la valeur *p* ajustée par une correction de Bonferroni ainsi que le taux de réussite des réponses.

de /y/, dont les valeurs F2 se rapprochent fortement de celles du /u:/ anglais. Au contraire, un son de la L2 considéré comme « nouveau », tel que /u/ dans la paire /y/-/u/, serait plus facile à acquérir.

# Références


Abry D., Chalaron M-L. (2011). *Les 500 exercices de phonétique B1/B2 : avec corrigés*. Vanves : Hachette Français Langue Étrangère.

Akyüz A., Bazelle-Shahmaei B., Bonenfant J., Flament M-F., Lacroix J., Moriot D., Renaudineau P. (2002). *Exercices d'oral en contexte. Niveau débutant*. Paris : Hachette Livre Français langue étrangère.

Amand M., Touhami Z. (2015). "Righ' here, righ' now." Immediate pronunciation versus audio and visual corrections in second-language speech: unreleased plosives by French learners of English. Actes de *EUROSLA 2015*, 77.

Audacity development team (2008). Audacity (Version 2.0.5. [Computer software]. Accessible : audacity.sourceforge.net/download). http://audacity.fr/. (12/ 2016)

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Best C., Faber A., Levitt A. (1996). Assimilation of non-native vowel contrasts to the American English vowel system. *Journal of The Acoustical Society of America* 99. 2602-2603.

Birdsong D . (2003). Authenticité de prononciation en français L2 chez des apprenants tardifs anglophones : analyses segmentales et globales. *Acquisition et interaction en langue étrangère* 18, 17-36.

Boersma P . , Weenink D . (2017). Praat: doing phonetics by computer [logiciel]. http://www.praat.org/ (02/2017)

Colantoni L., Steele J., Escudero P. (2015). *Second language speech: Theory and practice*. Cambridge, Royaume-Uni : Cambridge University Press.

Escudero P. (2007). Second-language phonology: The role of perception. Dans Pennington M.C. (dir), *Phonology in context* (p. 109-134). Basingstoke, Royaume-Uni : Palgrave Macmillan.

Flege J. E. (2002). Interactions between the Native and Second-language Phonetic Systems. In Burmeister P., Piske T., et Rohde A. (dir), *An Integrated View of Language Development: Papers in Honor of Henning Wode* (p. 217-244). Allemagne : Wissenschaftlicher Verlag Trier.

Gendrot C. (s.d.). My Praat and other scripts. Repéré à from http://gendrot.ilpga.fr/scripts.htm

Hothorn T., Hornik K., Zeileis A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* 15(3), 651– 674.

Levy E. S. (2009). Language experience and consonantal context effects on perceptual assimilation of French vowels by American-English learners of French. *The Journal of the Acoustical Society of America* 125(2), 1138-1152.

Major R. C. (2001). *Foreign accent: The ontogeny and phylogeny of second language phonology*. Oxford : Routledge.

Mayr R., Escudero P. (2010). Explaining individual variation in L2 perception: Rounded vowels in English learners of German. *Bilingualism: Language and Cognition* 13. 279-297.

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/. (02/2017)

# Perturbation de l'organisation temporelle de la parole suite à un effort physique

[1]Camille Fauth, [1]Angéline Duchemin, [1]Béatrice Vaxelaire, Rudolph Sock[1,2]
[1]Université de Strasbourg, Institut de Phonétique, E.A. 1339 LiLPa, Strasbourg, France
[2]Language Information and Communication Laboratory (LICOLAB) Université Pavla Jozefa Šafárika (UPJŠ) / Košice (Slovaquie)
cfauth@unistra.fr

## RESUME

La production de la parole, dans son décours temporel, est organisée en groupes rythmiques conditionnés par des contraintes physiologiques, linguistiques et communicatives. L'objectif de ce travail est d'évaluer les conséquences d'un effort physique sur cette organisation temporelle de la parole, en se focalisant sur les pauses (respiratoires et silencieuses) produites lors d'une tâche de lecture. 14 locuteurs répartis en trois groupes (fumeur, contrôle et sportifs) ont été enregistrés avant et après avoir fourni un effort physique afin d'observer si leur condition physique pouvait avoir un effet sur leurs productions. Les résultats indiquent que les locuteurs réorganisent la réalisation des pauses. Ainsi, si le nombre global de pauses réalisées reste identique entre les deux phases d'enregistrement, ce sont les pauses silencieuses qui sont les plus nombreuses avant l'effort, alors que ce sont les pauses respiratoires qui sont les plus importantes après l'effort, et ce pour tous les groupes de locuteurs. Les vitesses d'élocution et d'articulation restent inchangées.

## ABSTRACT

Speech production, in its temporal course, is organised into rhythmic groups conditioned by physiological, linguistic and communicative constraints. The main thrust of this work is to evaluate the consequences of physical effort on this temporal organisation of the speech, by focusing on pauses (breathy and silent pauses) produced during text reading. To do so, 14 speakers, divided into three groups (controls, smokers and athletes), were recorded before and after performing a series of skipping jumps, in order to see if their physical condition could have an effect on readjustment strategies deployed during the production of speech in a breathlessness condition. Results indicate that the way pauses are carried out by speakers is reorganised. Thus, if the total number of pauses observed remains the same across the two recording phases, silent pauses are the most numerous before effort, whereas breathy pauses are more significant after effort, such results being true for all groups of speakers. Speech and articulation rates remain unchanged.

MOTS-CLES : production de la parole ; timing ; perturbations ; efforts physiques

KEYWORDS: speech production, timing, perturbations, physical effort

# 1 Introduction

L'organisation des cycles respiratoires (*breath cycles*) lors de la production de la parole diffère selon différents paramètres tels que les stratégies individuelles (Teston & Autesserre, 1987), l'âge du locuteur (Sperry, Klich, 1992), la complexité de la planification (Mitchell et al., 1996), la nature


10.21437/JEP.2018-28

de la tâche de production (Wang, et al., 2010)… Le cycle respiratoire lors de la production de la parole est d'autant plus irrégulier qu'il repose sur des contraintes physiologiques, linguistiques et communicatives (Rochet-Capellan, Fuchs, 2013). Les contraintes physiologiques englobent le besoin de ventilation pour la production de la parole et la gestion de la respiration pour des raisons métaboliques. Les contraintes linguistiques sont déterminées par la structure grammaticale de la phrase et par l'adaptation de celle-ci au sein d'un ou de plusieurs cycle(s) respiratoire(s). Les contraintes communicatives sont celles résultant du besoin de structuration du discours à travers le groupe rythmique qui, lui, impose au cycle respiratoire une certaine organisation.

Les études sur la qualité vocale du fumeur sont nombreuses et l'on sait que fumer peut avoir un effet très néfaste sur la capacité respiratoire, mesurée à l'aide d'un spiromètre, notamment lorsque sa consommation a été commencée jeune (Urrutia et al., 2005). En revanche, l'exercice physique a un bienfait essentiel sur le système respiratoire, puisqu'il augmente la ventilation et la circulation sanguine dans les bronches et les poumons. Ce bienfait peut en effet affecter l'état général du métabolisme humain (Saltin, Grimby, 1968 ; Pate et al., 1995). Cependant, tout de suite après un effort physique, parler devient plus difficile et la respiration est saccadée. Ce trouble est provoqué par le besoin compétitif entre la respiration pour des raisons métaboliques et la respiration pour la production de la parole (Trouvain, Truong, 2015).

Les pauses jalonnent la production de la parole et permettent l'alternance entre des phases de production et des phases de silences. Ces derniers, sont considérés comme des pauses vides et sont alors définis comme une interruption du flux de parole se répercutant sur le signal acoustique par une amplitude nulle ou non-significative (Duez, 2003). Notons toutefois que d'un point de vue acoustique, lors de la production de la parole, les phases de silence sont rares puisque les locuteurs produisent souvent des clics, des bruits de déglutition ou des bruits de respiration. Les bruits de souffle dans les pauses peuvent d'ailleurs remplir plusieurs fonctions. D'un point de vue syntaxique par exemple, ils peuvent signaler la longueur de la phrase à venir ou marquer une pause de niveau supérieur (Grosjean, Collins, 1979 ; Strangert, 1991 ; Fuchs et al, 2013). Ils peuvent même avoir un intérêt pour la mémorisation des phrases (Whalen, et al. 1995). En outre, une modification de la configuration des voies respiratoires, par exemple en induisant un stress physique, peut avoir un impact important sur la structure de la phrase prosodique (Trouvain, Truong, 2015). Enfin, habituellement les pauses silencieuses en fin de phrases sont considérées comme une interruption adéquate du flux articulatoire, tandis que les pauses apparaissant à d'autres endroits sont considérées comme des disfluences, comparables aux pauses remplies par « euh/m ». La question du seuil à partir duquel il est possible de considérer une pause silencieuse varie par exemple de 100ms (Trouvain, 2004) à 200ms (Lennon, 1990 ; Cucchiarini et al. , 2002) voire à 400ms (Tavakoli, 2011).

L'objectif de ce travail est d'étudier l'organisation temporelle de la lecture d'un texte après un effort physique, produit par trois groupes de locuteurs aux conditions physiques différentes (fumeurs, contrôles, sportifs). Nous souhaitons savoir si la condition physique pourrait avoir un effet sur les stratégies de réajustement que les locuteurs peuvent déployer face à l'essoufflement, ou la perturbation du cycle respiratoire, durant la production de la parole.

# 2 Méthodologie

## 2.1 Corpus et participants

Les sujets ont été enregistrés dans la chambre anéchoïque de l'Institut de Phonétique de Strasbourg, en position debout, à 20 cm d'un microphone unidirectionnel, relié à un enregistreur numérique. Les emplacements du sujet et du microphone ont été matérialisés au sol pour limiter au maximum les déplacements entre les différentes sessions d'enregistrement. Les locuteurs avaient pour tâche de lire à une hauteur et à une vitesse confortables le texte *la bise et le soleil* après avoir effectué deux minutes consécutives de corde à sauter. Préalablement, une lecture de référence, au repos, avait été effectuée. Ils avaient également eu un certain temps pour prendre connaissance du texte.

Pour cette étude, 14 sujets, 9 hommes et 5 femmes, âgés de 25 ans en moyenne (E.T. 4,59) ont été retenus. Ils avaient tous pour langue maternelle le français, ne présentaient aucune pathologie vocale, respiratoire ou auditive. De plus, compte tenu de l'effort physique qui leur est demandé, leur IMC (Indice de Masse Corporel) et leur indice d'Ashwell (ratio tour de taille/hauteur, et l'état cardio-métabolique) ont également été contrôlés. Ces locuteurs ont été répartis en trois groupes. Le premier groupe est constitué de 6 locuteurs « contrôle » non-fumeurs qui ne pratiquent pas d'activité physique régulière. Le deuxième groupe est composé de 4 locuteurs qui n'ont pas d'activité sportive régulière et qui fument plus de 5 cigarettes par jour depuis plus de deux ans. Le troisième groupe est formé par 4 sujets pratiquant un sport aquatique au moins deux fois par semaine depuis deux ans.

## 2.2 Mesures

Les données acoustiques ont été analysées à l'aide de Praat (Boersma, Weenink, 2016). Elles ont été segmentées de façon semi-automatique grâce à EasyAlign (Goldman, 2011). La détection des pauses s'est faite à partir d'indices perceptivo-visuels, en d'autres termes, les pauses silencieuses correspondent à un silence perceptible, accompagné d'une rupture d'activité acoustique visible sur le signal de parole, tandis que les pauses respiratoires sont repérées à partir d'une prise de souffle ou d'une expiration audible et visible (amplitude non nulle). Aucun seuil de durée n'a été appliqué. Les pauses initiales et finales n'ont pas été prises en compte. La tenue des occlusives non voisées, précédée d'une pause, a été fixée à 50ms.

Nous avons quantifié pour chaque tâche et chaque locuteur : le nombre et la durée des pauses silencieuses (PS) et des pauses respiratoires (PR). Une pause respiratoire a été déterminée à l'aide d'indices perceptivo-visuel et a été considérée comme un élément unique à partir du moment où un bruit de souffle était audible et visible. De plus ont été calculées : la vitesse d'élocution (VE), c'est-à-dire le nombre de syllabes par seconde en prenant en compte la durée des pauses, et la vitesse d'articulation (VA) ou le nombre de syllabes par secondes sans tenir compte de la durée des pauses.

## 2.3 Hypothèses

Ce corpus annoté devrait nous permettre d'étudier les stratégies de lecture que différents locuteurs adoptent lorsqu'ils sont soumis à un effort physique. Nous pouvons supposer que, compte tenu de l'effort physique fourni, induisant une compétition entre la respiration pour des raisons métaboliques et la respiration pour la production de la parole, l'ensemble des locuteurs pourrait produire des

pauses (silencieuses et respiratoires) plus nombreuses et plus longues lors de la deuxième phase d'enregistrement (1). Parmi les pauses réalisées, ce sont les pauses respiratoires qui devraient être significativement plus longues et plus nombreuses après l'effort (2). En ce qui concerne les différents groupes de locuteurs, les différences de l'organisation temporelle de la parole entre les deux conditions de production devraient être plus importantes chez les fumeurs, puisqu'ils devraient être plus sujets à un essoufflement remarquable. Corolairement, les performances seraient moins modifiées chez les locuteurs sportifs (3). Compte tenu de ces modifications, la vitesse d'élocution deviendrait plus lente (soit moins de syllabes par seconde) lors de l'enregistrement post effort physique, la vitesse d'articulation pourrait être augmentée (soit plus de syllabes par seconde), les locuteurs cherchant néanmoins à optimiser leur vitesse de production de la parole, ainsi que l'intelligibilité de celle-ci, sous la pression de l'essoufflement (4).

# 3 Résultats

Les statistiques ont été réalisées grâce au logiciel RStudio (2015). Le seuil de significativité a été considéré comme suit : $p < 0,05$. Les comparaisons de la distribution des différentes pauses ont été effectuées à l'aide d'un test de Kolmogorov-Smirnov, tandis que les comparaisons de durées ont été réalisées avec un U-test de Mann-Whitney-Wilcoxon.

## 3.1 Pauses respiratoires (PR) et pauses silencieuses (PS)

### *3.1.1 Toutes pauses confondues*

Dans un premier temps, nous avons étudié la distribution de toutes les pauses (silencieuses et respiratoires) confondues, produites par nos locuteurs pendant la lecture du texte. Ainsi, nous avons compté la réalisation de 306 pauses (soit 21,26 pauses en moyenne, E.T. 6,53) avant l'effort et de 299 pauses après l'effort (soit 21,85 pauses en moyenne, E.T. 5,46) ; cette différence n'est naturellement pas significative.

En revanche, si l'on considère la durée de ces pauses (respiratoires et silencieuses) produites par l'ensemble des locuteurs après l'effort, nous constatons qu'elles sont significativement ($p = 0.000$) plus longues (638ms en moyenne E.T. 373) que celles réalisées avant l'effort (448ms en moyenne E.T. 474). Les écarts-types sont importants pour les deux phases d'enregistrement, probablement parce que nous n'avons pas appliqué de seuil de détection des pauses (voir FIGURE 1). Ce paramètre ne permet toutefois pas de distinguer entre nos différents groupes de locuteurs ($p = 0.494$), si les fumeurs ont effectivement des pauses plus longues après l'effort, les productions des locuteurs « contrôle » et sportifs se rapprochant.

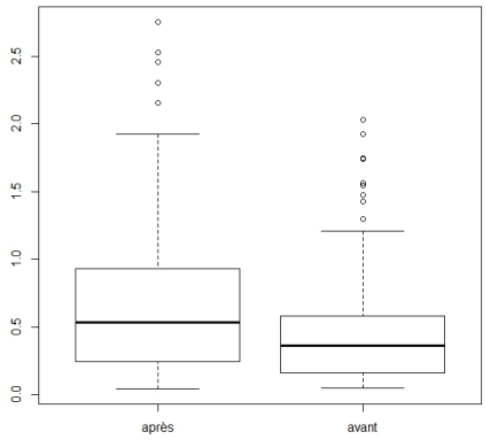



FIGURE 1: Durées des pauses en secondes (PS et PR) pour tous les locuteurs après et avant effort

### *3.1.2 Pauses respiratoires*

Dans un deuxième temps, nous avons cherché à savoir si les locuteurs réalisaient plus de pauses respiratoires lorsqu'ils sont essoufflés. Quantitativement, ils réalisent effectivement plus de pauses respiratoires (222 pauses, soit une moyenne de 15,86 pauses, E.T. 7,09, par locuteur) après l'effort qu'avant l'effort (130 pauses, soit une moyenne de 9,28 pauses, E.T. 2,81, par locuteur). Notons que ces différences sont significatives ($p = 0.007$). Comme précédemment, les pauses respiratoires produites après l'effort sont significativement ($p = 0.013$) plus longues pour tous les locuteurs (voir FIGURE 2). Ainsi, elles durent en moyenne 619ms (E.T. 309) avant l'effort et 796ms (E.T. 457) après l'effort. Ce paramètre ne permet toutefois pas de distinguer entre les différents groupes de locuteurs ($p = 0.348$). Notons toutefois (voir FIGURE 4) qu'après l'effort, ce sont les locuteurs sportifs qui réalisent en moyenne les pauses respiratoires les plus longues.

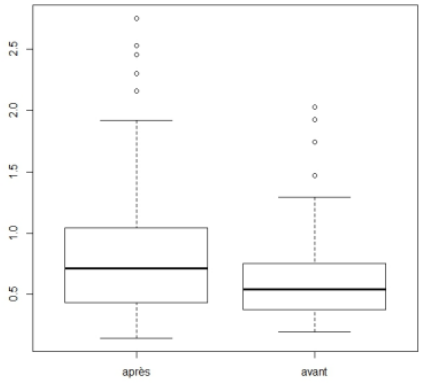



FIGURE 2 : Durées des pauses respiratoires en secondes pour tous les locuteurs après et avant effort

### *3.1.3 Pauses silencieuses*

Dans un troisième temps, nous avons investigué la réalisation des pauses silencieuses. De façon intéressante, c'est le schéma inverse à celui présenté précédemment qui se réalise, c'est-à-dire qu'après l'effort, les locuteurs ne réalisent que 77 pauses (soit 5,5 pauses en moyenne, E.T. 3,55, par locuteur) alors qu'ils en réalisaient 176 (soit 12,57 pauses en moyenne, E.T. 5,89, par locuteur) avant l'effort. Cette différence est significative ($p=0.001$). Parallèlement (voir FIGURE 3), la durée des pauses avant et après l'effort est également significativement différente ($p = 0.007$) si l'on

considère l'ensemble des locuteurs. Ainsi, elles durent en moyenne 209 ms (E.T. 237) après l'effort, mais 288 ms (E.T. 296) avant l'effort.

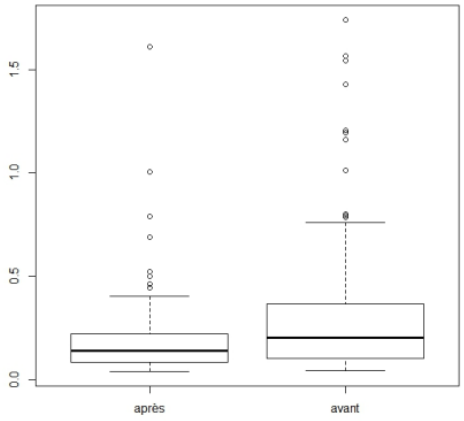



FIGURE 3 : Durées des pauses silencieuses en secondes pour tous les locuteurs après et avant effort

### *3.1.4 Ratio du nombre de pauses respiratoires et silencieuses*

Compte tenu des résultats précédents, il nous semble intéressant de les synthétiser en proposant de présenter pour l'ensemble des locuteurs, la distribution des pauses respiratoiers et des pauses silencieuses, lors des deux phases d'enregistrment (voir FIGURE 4). Il apparait ainsi de façon relativement clair qu'entre les deux phases d'enregistrement, les locuteurs réalisent un nombre de pauses relativement constants mais dépendant du locuteur, et ne font que les redistribuer entre les pauses silencieuses (en rouge, plus nombreuses avant l'effort, sur le graphique de gauche) et les pauses respiratoires (en bleu, plus nombreuses après l'effort, sur le graphique de droite).

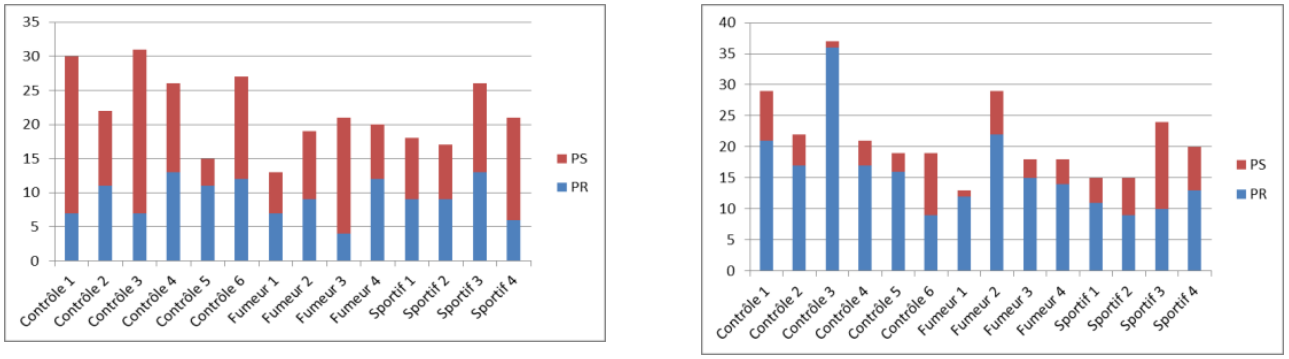



FIGURE 4 : Ratio du nombre de pauses silencieuses (PS en rouge) et de pauses respiratoires (PR en bleu) pour tous les locuteurs avant effort (à gauche) et après effort (à droite)

### *3.1.5 Variabilité inter-individuelle*

LA FIGURE 5 montre que la durée des pauses (silencieuses et respiratoires) est soumise à beaucoup de variations inter-individuelles. Si certains locuteurs, comme le locuteur contrôle 2 par exemple, maintiennent des durées comparables et peu variables avant et après l'effort, d'autres présentent de grandes différences, comme le locuteur fumeur 2, par exemple, chez qui la durée et la variabilité des pauses augmentent après l'effort. De façon générale toutefois, la variabilité est moins importante avant l'effort (données en bleu) et les pauses généralement plus courtes. En revanche, la condition physique des locuteurs ne permet pas de prédire leurs stratégies post-effort.

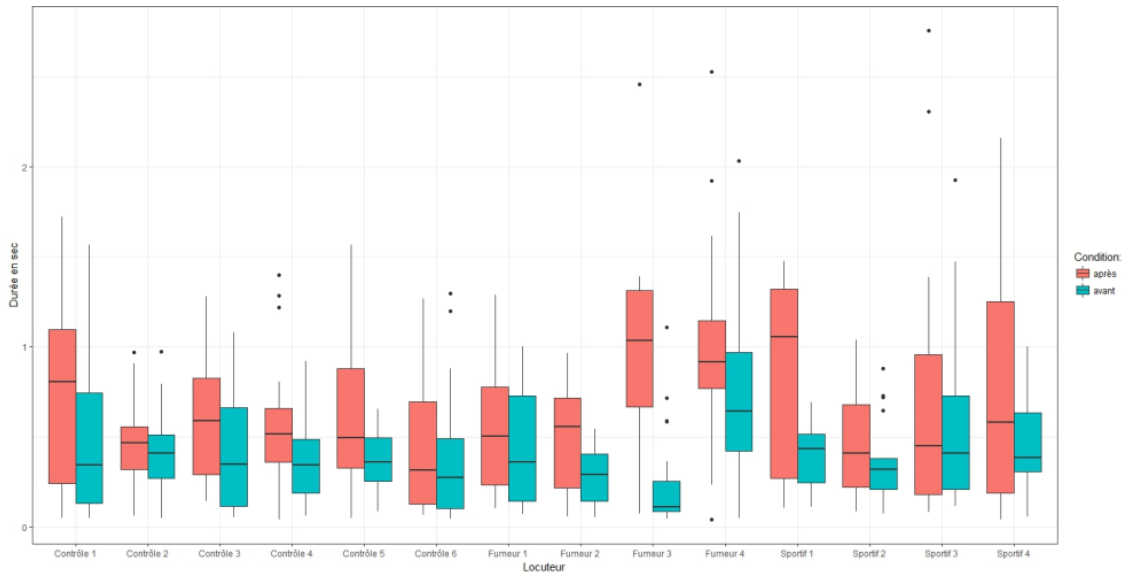



FIGURE 5 : Durées des pauses (PR et PS) en fonction de la tâche pour tous les locuteurs

### 3.2 Vitesses d'élocution et d'articulation

Enfin en dernier lieu, nous avons cherché à quantifier les vitesses d'élocution et d'articulation en syllabes par seconde (en prenant en compte les pauses ou sans prendre en compte les pauses, respectivement). En ce qui concerne ces deux mesures, elles ne sont pas modifiées par l'effort physique. La vitesse d'articulation est, en moyenne, de 6,15 syllabes par seconde (E.T. 0,75) avant l'effort et de 6,25 syllabes par seconde (E.T. 0,65) après l'effort. La vitesse d'élocution, elle, est en moyenne de 4,71 syllabes par seconde (E.T. 0,73) avant l'effort et de 4,26 syllabes par seconde après l'effort (E.T. 0,61). La condition physique du locuteur n'influence pas ce paramètre, les valeurs ne sont pas significativement différentes en fonction des différents groupes de locuteurs.

## 4 Conclusions

Dans ce travail, nous avons cherché à évaluer les conséquences d'un effort physique sur les stratégies de lecture chez 14 locuteurs à la condition physique différente, répartis dans 3 groupes (contrôle, fumeur et sportif). Notre première hypothèse n'est que partiellement confirmée puisque si les locuteurs produisent effectivement des pauses (respiratoires et silencieuses) plus longues après l'effort, elles ne sont pas pour autant plus nombreuses. En revanche, si l'on ne considère que les pauses respiratoires (avec prise de respiration visible sur le signal acoustique), celles-ci sont significativement plus longues pour l'ensemble des locuteurs après l'effort de même que leur nombre est significativement augmenté entre les deux enregistrements. L'étude des pauses silencieuses a permis de montrer un schéma opposé, à savoir que les pauses silencieuses après l'effort sont significativement moins nombreuses et plus courtes pour l'ensemble des locuteurs. En ce qui concerne les trois groupes de locuteurs, nos mesures n'ont pas permis de mettre au jour des stratégies particulières à un groupe de locuteurs. Les locuteurs fumeurs ont effectivement des performances moins bonnes que les autres locuteurs, sans que cela soit significatif. En revanche, les locuteurs sportifs et ceux qui ne pratiquent pas d'activité sportive régulière ont des résultats relativement proches. Notons qu'une pause a été considérée comme respiratoire à partir du moment

où il y avait une activité respiratoire observable sur le signal acoustique, ces résultats pourront être affinés en segmentant finement les pauses, en fonction des différents événements (expiration, silence, inspiration) qui peuvent se réaliser à l'intérieur des pauses que nous avons considérées comme un seul élément. Enfin, les vitesses d'articulation et d'élocution ne sont pas modifiées par l'effort physique. Les locuteurs réorganisent le ratio pauses respiratoires/pauses silencieuses après l'effort sans toutefois optimiser leur vitesse de production de la parole. Nous pouvons supposer que ces « choix » sont contraints par les besoins d'intelligibilité, avec notamment la nécessité de structurer les énoncés en groupe rythmiques, ou en groupes de sens, qui se rapprochent de ceux généralement observés en production plus ou moins naturelle de la parole. Enfin, il convient de signaler que la variabilité inter-individuelle est un paramètre important. Conformément aux travaux antérieurs, notre étude confirme que la lecture d'un même texte, avec la même consigne, peut conduire à des productions de nature très différentes et que les choix stratégiques de lecture se répercutent notamment sur la réalisation des pauses, et ce indépendamment de la condition physique du locuteur. Ces résultats sont notamment à prendre en compte dans les protocoles de rééducation des pathologies qui affectent le système pneumo-phonatoire.

# 5 Perspectives

Ce travail pourra être enrichi, d'une part, en augmentant le nombre de locuteurs dans les différents groupes et, d'autre part, en ajoutant d'autres catégories de locuteurs tels que les instrumentistes à vent ou les professionnels de la parole.

Dans ce travail, les pauses ont été étudiées en termes de distribution et de durée, il pourrait être intéressant d'observer l'endroit où elles se réalisent. De plus, d'autres paramètres tels que la qualité vocale, d'autres éléments prosodiques ou l'intensité des prises de respiration pourraient également être ajoutés. Il nous semble aussi pertinent d'étudier plus en détails les pauses respiratoires pour voir qu'elle est la proportion de prise de souffle dans ces éléments. Ces résultats pourraient également être comparés à des données en parole spontanée, puisque les stratégies des locuteurs dépendent en partie de la tâche de parole (Campione & Véronis, 2004). Enfin, il nous semble que ce travail est un bon point de départ pour conduire des investigations sur la gestion des flux respiratoires (thoracique et abdominal) à l'aide d'un système d'acquisition de telles données, comme Respitrace®.

# Remerciements

Les auteurs remercient Clémence Duroux et Solange Martay, étudiantes à l'école d'Orthophonie de Strasbourg, qui ont activement participé à l'élaboration de ce travail dans le cadre de leur stage recherche.

# Références

BOERSMA P., WEENINK D. (2016). Praat: doing phonetics by computer (Version 6.0.21).

CAMPIONE E., VERONIS J. (2004). Pauses et hésitations en français spontané. Présenté à XXVème Journées d'Etude sur la Parole, Fès - Maroc.

DUEZ D. (2003), « Le pouvoir du silence et le silence du pouvoir : comment interpréter le discours politique », *MediaMorphoses*, (8), 77‑82.

FUCHS S., PETRONE C., KRIVOKAPIĆ J., HOOLE P. (2013). Acoustic and respiratory evidence for utterance planning in German. *Journal of Phonetics*, *41*(1), 29‑47.

GOLDMAN J.-P. (2011). EasyAlign: an automatic phonetic alignment tool under Praat. In *Interspeech 2011* (p. 3233‑3236). Florence.

GROSJEAN F., COLLINS M. (1979). Breathing, pausing and reading. *Phonetica*, *36*(2), 98‑114.

MITCHELL H L., HOIT J D., WATSON P J. (1996). Cognitive-linguistic demands and speech breathing. Journal of Speech and Hearing Research, 39(1), 93‑104.

PATE R R, PRATT M, BLAIR, S N , HASKELL W L., MACERA C. A., BOUCHARD C., … KING A. C. (1995). Physical activity and public health. A recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine. *JAMA*, *273*(5), 402‑407.

ROCHET-CAPELLAN A., FUCHS S. (2013). The interplay of linguistic structure and breathing in German spontaneous speech. In *14th Annual Conference of the International Speech Communication Association (Interspeech 2013)* (p. 1228). Lyon, France.

SALTIN B, GRIMBY G. (1968). Physiological analysis of middle-aged and old former athletes. Comparison with still active athletes of the same ages. *Circulation*, *38*(6), 1104‑1115.

SPERRY E E., KLICH R J. (1992). Speech breathing in senescent and younger women during oral reading. *Journal of Speech and Hearing Research*, *35*(6), 1246‑1255.

STRANGERT E. (1991). Pausing in texts read aloud. In *Procedings of International Congres of Phonetics Science (ICPHS)* (Vol. 4, p. 238‑241). Aix-en-Provence (France).

TESTON B., AUTESSERRE D. (1987). L' aérodynamique du souffle phonatoire utilisé dans la lecture d'un texte en français (p. 33‑36). Présenté à International Congress of Phonetic Sciences (ICPhS), Tallin, Estonie.

TROUVAIN J., TRUONG K P. (2015). Prosodic characteristics of read speech before and after treadmill running. In *Interspeech 2015*. Dresde (Allemagne): International Speech Communication Association (ISCA).

URRUTIA I., CAPELASTEGUI A., QUINTANA J M., MUÑIOZGUREN N., BASAGANA X., SUNYER J., ECRHS-I. (2005). Smoking habit, respiratory symptoms and lung function in young adults. *European Journal of Public Health*, *15*(2), 160‑165.

WHALEN D. H., HOEQUIST C. E ., SHEFFERT S. M. (1995), « The effects of breath sounds on the perception of synthetic speech », *The Journal of the Acoustical Society of America*, *97*, 3147‑3153.

WANG Y T., GREEN J R., NIP I S B., KENT R D., KENT J F. (2010). Breath Group Analysis for Reading and Spontaneous Speech in Healthy Adults. *Folia Phoniatrica et Logopaedica*, *62*(6), 297‑302.

XXXIIe Journées d'Études sur la Parole
4-8 juin 2018, Aix-en-Provence, France

# Prédiction *a priori* de la qualité de la transcription automatique de la parole bruitée

Sébastien Ferreira[1, 2] Jérôme Farinas[1] Julien Pinquier[1] Stéphane Rabant[2]
(1) IRIT, Université de Toulouse, CNRS, Toulouse, France
(2) Authôt, 52 Avenue Pierre Semard, 94200, Ivry-sur-Seine, France
prenom.nom@irit.fr[1],sferreira@authot.com,srabant@authot.com

RÉSUMÉ
De nombreuses sources de variabilité dégradent les performances d'un système de Reconnaissance Automatique de la Parole (RAP). Dans cette étude, les dégradations provoquées par le type et le niveau de bruit sont explorées afin de prédire *a priori* la qualité de la RAP, i.e. avant même le décodage. Notre méthode se fonde sur une séparation spectrale de la parole et du bruit afin de produire un modèle de régression. L'expérimentation a été réalisée sur le corpus Wall Street Journal, bruité avec le corpus NOISEX-92 (17 types de bruit) que nous appliquons à 9 niveaux de rapport signal à bruit. La méthode de régression proposée obtient moins de 8% d'erreur moyenne entre le Word Error Rate (WER) prédit et le WER réellement obtenu par le système de transcription automatique de la parole.

ABSTRACT
**A priori prediction of the quality of automatic speech to text conversion for noised speech.**

Many sources of variability degrade the performance of Automatic Speech Recognition (ASR) system. In this study, the degradations caused by the type and level of noise are explored in order to predict the *a priori* quality of ASR, i.e. even before decoding. Our method is based on a spectral separation of speech and noise to produce a regression model. The experiment was carried out on the Wall Street Journal corpus, noised with the NOISEX-92 corpus (17 types of noise) that we apply at 9 levels of signal-to-noise ratio. The proposed regression method obtains less than 8% of mean error between the predicted Word Error Rate (WER) and the actual WER obtained by automatic speech to text conversion system.

MOTS-CLÉS : prédiction d'erreur, reconnaissance automatique de la parole, analyse du bruit, séparation de la parole et du bruit.

KEYWORDS: error prediction, automatic speech recognition, noise analysis, speech/noise discrimination.

## 1 Introduction

Les progrès effectués dans le domaine de la Reconnaissance Automatique de la Parole (RAP) permettent de transcrire un fichier audio en texte dans des situations de plus en plus complexe. Les performances obtenues par ces systèmes sont fortements liées aux méthodes et aux données utilisées pour l'apprentissage des modèles acoustiques et linguistiques. Actuellement, il n'existe pas de système de RAP qui serait efficace pour toutes les situations, car il existe de nombreuses sources de variabilité dans un signal de parole : l'environnement acoustique, la voix du locuteur, le manière


10.21437/JEP.2018-29

de parler, l'interaction entre les locuteurs, la thématique du discours... Pour réduire l'impact de ces différentes sources de variabilité, il est courant d'utiliser des systèmes spécifiques pour chaque cas d'utilisation : la dictée vocale, la commande vocale, les enregistrements télévisuels et radiophoniques, les réunions, l'enseignement, l'automobile...

Pour sélectionner le meilleur système de RAP, sans information préalable sur le fichier traité, il serait intéressant de pouvoir prédire *a priori* la qualité des transcriptions. Le Word Error Rate (WER) est une métrique couramment utilisée pour évaluer la qualité de la transcription. Il existe de nombreuses méthodes pour prédire le WER, mais, ces méthodes utilisent des informations qui dépendent des résultats d'un système de RAP : mesures de confiance (CM) (Jiang, 2005; Ghannay *et al.*, 2015), probabilités *a posteriori*, paramètres lexicaux et syntaxiques, posteriogram sur les phonèmes (Meyer *et al.*, 2017) ou diverses statistiques calculées sur les données d'entraînement (Hermansky *et al.*, 2013).

Dans cet étude, nous cherchons une méthode qui n'exige pas de score interne du système de RAP et qui puisse être calculée avant d'utiliser le système lui-même. En effet, la prédiction du WER obtenue par un système de RAP sur un fichier audio devra être déterminée *a priori*. Il existe de nombreuses sources d'erreur de reconnaissance possibles. Ces erreurs sont causées en particulier par l'environnement sonore, les dialectes sous-représentés, les abréviations, les noms propres, les voix atypiques, une mauvaise maîtrise du langage, une thématique trop spécifique... Afin de ne pas mélanger toutes les sources d'erreur possibles, cette première étude se concentre sur les dégradations de la parole uniquement causées par le bruit. Pour représenter un grand nombre de bruits ambiants, nous avons artificiellement bruité les données pour différents types de bruits et à différents niveaux de RSB (Rapport Signal sur Bruit), d'une manière similaire à l'expérience de Haitian Xu (Xu *et al.*, 2007) pour évaluer la prédiction *a priori* obtenue par notre méthode.

Tout d'abord, le système de prédiction du WER est présenté dans la section 2. Ensuite, le cadre expérimental est décrit dans la section 3. Puis, les résultats obtenus sont présentés dans la section 4.

# 2 Système d'estimation du WER

Les modèles acoustiques utilisés par les systèmes de RAP modélisent généralement des paramètres plus ou moins complexes liés à l'énergie à court terme du signal : par exemple des paramètres issus d'un spectrogramme. Or, comme le bruit provoque une perturbation de l'énergie à court terme, les paramètres utilisés par les modèles acoustiques, même s'il sont plus robustes, se retrouvent dégradés. De plus, lorsque le bruit devient important, i.e. losrque le RSB est très faible (voir négatif), le bruit recouvre la parole. Il est devient alors très difficile de distinguer l'énergie provenant de la parole de celle provenant du bruit. Une analyse de ce recouvrement a été faite dans de nombreux domaines comme l'estimation du RSB (nis, 1994), la détection d'activité vocale (Voice Activity Detection - VAD) (Yiming & Rui, 2015) et l'amélioration de la parole (Speech Enhancement - SE) (Ruwei *et al.*, 2016). Afin de quantifier l'impact du bruit sur un signal, il est courant d'estimer le RSB. Cependant, le RSB ne permet pas, à lui seul, de quantifier l'impact sur la qualité de la transcription des systèmes de RAP. Le type de bruit, la localisation du signal de parole en temps et en fréquence et la robustesse au bruit du système de RAP utilisé sont aussi à prendre en compte (figure 1). Nous proposons d'étudier le comportement de l'énergie à court terme du signal tout en tenant compte des différents facteurs cités précédement.

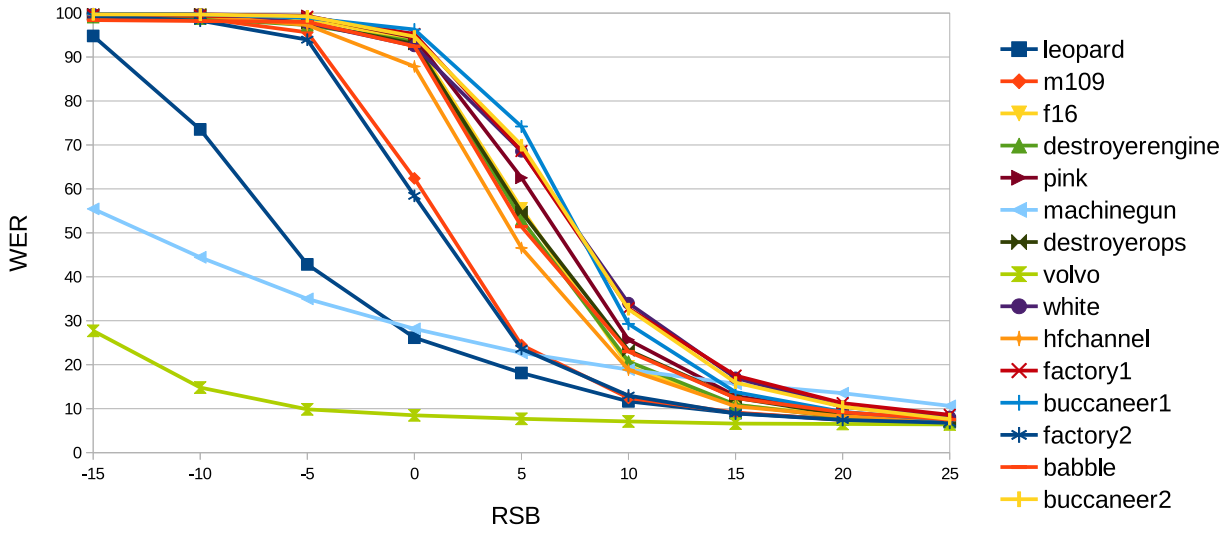



FIGURE 1 – Erreurs obtenues par un système entraîné dans des conditions propres en fonction du type de bruit et du RSB.

## 2.1 Architecture globale

Le système est composé de 4 étapes (voir figure 2) :

1. Calcul de la Transformée de Fourier Discrète (TFD) avec un fenêtrage de 512 points et un recouvrement de moitié puis normalisée sous l'échelle [0 ;1].
2. Détermination d'un masque binaire pour séparer la parole du bruit (voir section 2.2).
3. Extraction de paramètres pour chaque bande (voir section 2.3).
4. Régression entre les paramètres et les WER issus de l'apprentissage (voir section 2.4).

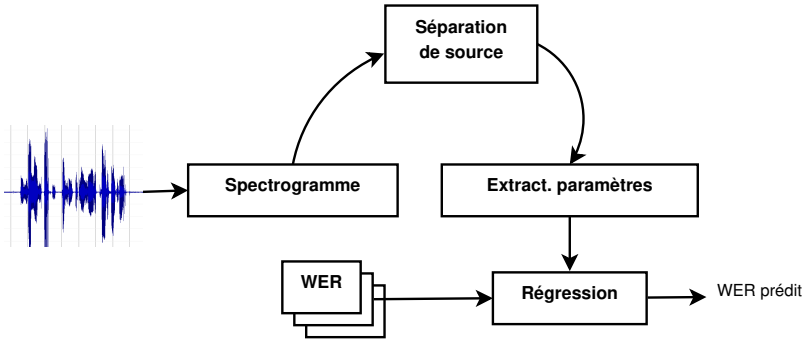



FIGURE 2 – Architecture du système de prédiction du WER.

## 2.2 Séparation de la parole et du bruit

Le bruit, en recouvrant le signal de parole, provoque trois types d'erreurs de reconnaissance :

— suppression : un phonème n'est pas reconnu, alors qu'il existe.

— insertion : un phonème est reconnu, alors qu'il n'est pas présent.
— substitution : un phonème erroné est reconnu.

Afin d'étudier la superposition de la parole et du bruit, une séparation est effectuée. Parmi les multiples méthodes de séparation de sources, il existe la méthode des masques. Le calcul d'un masque (binaire ou pondéré) permet de séléctionner les energies à court terme provenant de la parole. Dans ce domaine, le masque binaire idéal (Ideal Binary Mask - IBM) correspond au masque qui permet de séléctionner uniquement la parole. Il est généralement calculé pour un signal enregistré dans des conditions non bruitées. Le calcul automatique d'un masque binaire (Binary Mask - BM), le plus proche possible de l'IBM, est un enjeu majeur pour le domaine de l'analyse de scènes acoustiques (Wang, 2005). Actuellement, la détermination du BM qui minimise l'écart avec l'IBM se fonde sur une mesure de RSB local.

Dans notre méthode nous calculons un BM pour chaque bande Bark (Zwicker, 1961) du spectrogramme afin d'étudier le comportement de l'énergie à court terme de la parole et du bruit séparément :

$$BM(f,t) = \begin{cases} 1, & \text{si } E(f,t) \geq \omega * \overline{E(f)} \\ 0, & \text{sinon} \end{cases}$$

avec :

$$\overline{E(f)} = \frac{1}{t_{max}} * \sum_{t=1}^{t_{max}} E(f,t)$$

et $f$ la fréquence, $t$ la trame, $t_{max}$ le nombre total de trames, $E$ l'énergie à court terme et $\omega$ la pondération.

Il est important de noter que le seuil $T = \omega * \overline{E(f)}$ peut varier grandement en fonction de la bande Bark. Pour déterminer le BM optimal, on cherche le premier $\omega > \frac{max(\Delta densities)}{C}$ avec C une constante et $\omega_{optimal} > \omega_{maximum}$ (voir figure 3). Suite à la détermination du BM (deuxième image de la figure 4), nous éliminons les artefacts causés par les bruits résiduels : nous appliquons un masque local sur le BM afin de sélectionner que les zones ayant une densité supérieure à $\theta$. Le résultat de cette amélioration est visible dans la troisième partie de la figure 4. Pour plus de détails, vous pouvez écouter quelques résultats ici [1].

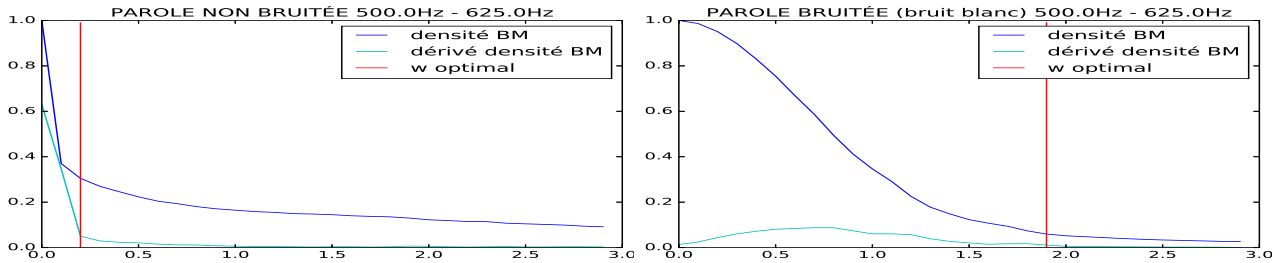



FIGURE 3 – Évolution de la densité du BM en fonction de $\omega$. À gauche le fichier non bruité, à droite le même fichier bruités avec du bruit blanc.

1. https://goo.gl/MAiXb1

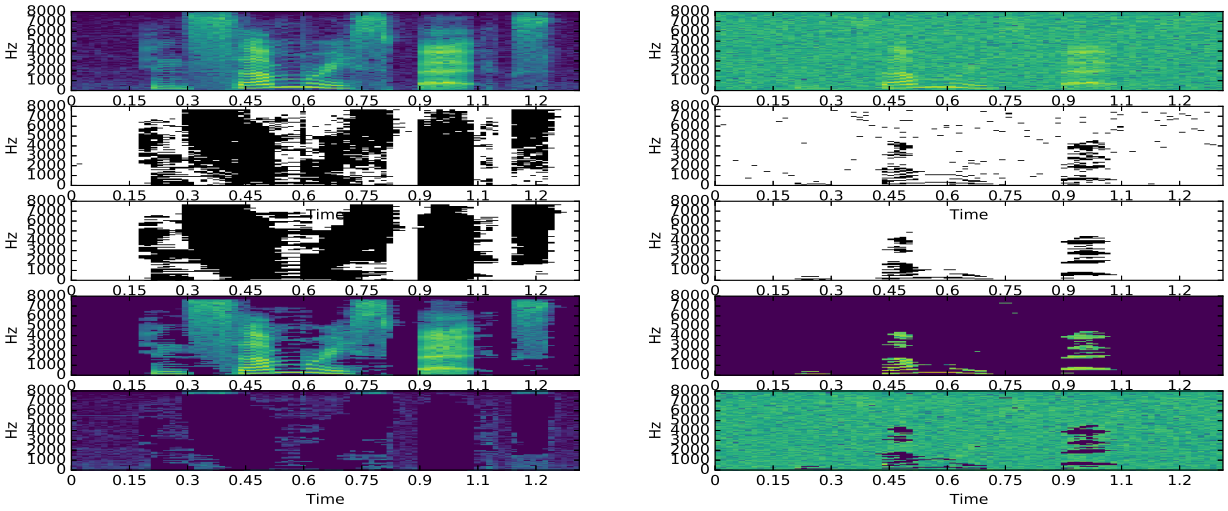



FIGURE 4 – Séparation de la parole et du bruit. À droite le fichier propre et à gauche le même fichier bruité avec du bruit blanc. De haut en bas : spectrogramme, BM, BM filtré, parole isolé, bruit isolé.

## 2.3 Extraction de paramètres

L'objectif est d'analyser le comportement de l'énergie à court terme du bruit et de la parole. Pour chaque bande Bark de 0 à 4400 Hertz le 5ème percentile, le 95e percentile et les 4 premiers moments (moyenne, variance, skewness et kurtosis) sont extraits. Une dernière mesure, similaire au Mean Crossing Rate (MCR) (Eyben, 2015), permet de quantifier le nombre de variations significatives de l'énergie à court terme. Ce paramètre est en quelque sorte un Zero Crossing Rate (ZCR) calculé sur la dérivée de l'énergie. Si les variations sont trop faibles (en dessous d'un seuil fixé à 0,2 dans notre cas) alors elles ne sont pas compatibilisées.

## 2.4 Régression

Un modèle de régression est calculé entre les paramètres extraits précédemment et le WER obtenu par différents systèmes de RAP. Cette regression permet de prendre en compte la variabilité des systèmes de RAP en fonction du volume et du type de bruit (Xu *et al.*, 2007) :

— les sytèmes multi-conditions sont plus performants sur des données bruitées,
— les sytèmes appris sur des données propres sont plus efficaces sur les données propres,
— les systèmes appris sur un type de bruit sont plus efficaces sur ce type de bruit.

Ces différences de performance proviennent de l'adéquation entre les données d'entrainement et les données de test. De plus, certains systèmes de RAP appliquent des algorithmes de SE en prétraitement, qui sont eux aussi plus ou moins efficaces selon le type de bruit. Deux types de modèle de régression ont été utilisés : la régression linéaire et la régression par Perceptron Multi-Couches (PMC). Scikit-learn (Pedregosa *et al.*, 2011) a été utilisé pour calculer les régressions. Notre MLP utilise 3 couches cachées et 18 (bandes Bark) * 6 neurones par couche.

# 3 Expériences

## 3.1 Corpus

Le corpus Wall Street Journal (WSJ) (John, DVD Philadelphia Linguistic Data Consortium 1993; wsj, Philadelphia Linguistic Data Consortium 1994), est fréquemment utilisé en RAP. Ce corpus a été choisie pour limiter les erreurs induites par le modèle de langage, les accents et les disfluences. Les sous-ensembles train_si284, dev93 et eval92 ont été séléctionnés lors de cette expérience. The details concerning the data size are indicated in the table 1.

Nous utilisons également le corpus NOISEX-92 (Varga & Steeneken, 1993) pour bruiter artificiellement nos données. Ce corpus est composé de 15 types de bruit d'une durée de 3min 56s chacun. La parole et le bruit ont été mixés pour neuf niveaux de RSB (de -15dB à 25dB par pas de 5dB) pour les 15 types de bruit (voir tableau 2). La fonction *v_addnoise* de la boîte à outils VoiceBox (Brookes, 2006) a été utilisée pour faire le mixage. La séléction de $x$ secondes de bruit pour bruiter un tour de parole est aléatoire. Suite au mixage, nous disposons d'un corpus final composé de 83h pour 136 conditions différentes : 15 (types de bruit) * 9 (niveaux de SNR) * $x$ (phrases sélectionnées dans la table 1) + $x$ (pour la condition parole propre).

Les données train_si284 sont utilisées pour entraîner les différents modèles acoustiques des systèmes de RAP. Les données dev93 et eval92 ont été séparées en deux nouveaux sous-ensembles pour entraîner et pour tester la régression.

TABLE 1 – Corpus WSJ utilisé.

| nom | nb locuteurs | nb phrases | temps |
|---|---|---|---|
| train_si284 | 284 | 37318 | 81h 15min |
| dev93 | 10 | 503 | 1h 5min |
| eval92 | 8 | 333 | 42min |
| Total | 302 | 38154 | 83h |

TABLE 2 – Types de bruit utilisés.

| nom | | |
|---|---|---|
| pink | factory1 | destroyerengine |
| f16 | babble | machinegun |
| white | leopard | destroyerops |
| m109 | factory2 | buccaneer1 |
| volvo | hfchannel | buccaneer2 |

## 3.2 Vérité terrain

Afin de réaliser une vérité terrain pour notre système de prédiction, différents systèmes de RAP ont été entrainés via Kaldi (Povey *et al.*, 2011) :

— un sytème pour des conditions propres,
— un système multi-conditions, entraîné avec les 15 types de bruit en fixant le RSB à 10dB,
— 15 systèmes mono-condition, en fixant le type de bruit et le RSB à 10dB.

Les systèmes de RAP ont été réalisés grâce à la recette de Karel Vesely [2] : DNN-HMM sur des triphones, vecteur de 40 dimensions (MFCC-LDA-MLLT-fMLLR), vocabulaire de 20k et le modèle de langage est un n-gram. Pour information, dans les conditions propres, notre système de RAP appris avec train_si284 obtient 5,84% de WER sur dev93 et 3,42% sur eval92.

Pour entraîner et tester la régression entre les paramètres extraits par notre méthode et le WER, un découpage des sous ensembles dev93 et d'eval92 a été effectué. Pour chaque locuteur, 60% des phrases provenant de dev93 et eval92 sont séléctionnées pour l'entraînement et 40% pour le test :

2. `http://kaldi-asr.org/doc/dnn1.html`

soit 1h 4min pour l'entraînement et 43min pour le test. Ce découpage a été effectué afin de limiter l'impact du locuteur.

# 4 Résultats

Pour déterminer le coefficient $\omega$ optimal pour le masque binaire (voir section 2.2), 30 masques binaires ont été testés en faisant varier $\omega$ de 0 à 3 par pas de 0,1. Ces 30 masques permettent de calculer l'évolution de la densités des BM. La constante $C$ a été fixée à 15 de manière empirique. Pour éliminer les valeurs aberrantes et ainsi améliorer le BM, la valeur de $\theta$ a été fixée à 0,4.

Afin d'évaluer la performance de la prédiction du WER (voir tableau 3), nous analysons l'erreur de prédiction absolu (PE pour Prediction Error) et l'écart type (SD pour Standard Deviation). La PE est calculée en moyennant les différences entre la prédiction et le WER réel pour chaque tour de parole. Le SD permet d'observer la dispersion de la prédiction à l'échelle d'un tour de parole. Dans la tableau 3, la PE et le SD sont affichés pour les deux régressions testées (linéaire et MLP). Le WER est prédit en utilisant différents systèmes de RAP pour évaluer l'indépendance de la méthode en fonction du système de RAP utilisée. Les résultats obtenus par le système de RAP ayant montré 5 différentes évolutions du WER en fonction du type de bruit, nous avons choisi d'afficher les résultats des 5 systèmes mono-condition correspondants. Les scores de PE obtenus sont généralement inférieurs à 8 sauf pour les systèmes mono-condition babble et factory. Par contre, la SD de la mesure reste importante (entre 10 et 11). Cette valeur indique que le WER ne peut pas être prédit efficacement pour une seule phrase : une fenêtre temporelle plus importante doit être utilisée. Cette même conclusion a été faite lors de l'expérience réalisé par Meyer (Meyer *et al.*, 2017), qui expliquait qu'un certain nombre de tours de parole devait être utilisé pour obtenir une prédiction sufisamment stable. Nous pouvons constater que la regression MLP obtient de meilleurs résultats que la régression linéaire.

TABLE 3 – Résultats des prédictions de WER pour différents systèmes de RAP.

| | | clean | multi-cond. | babble | fact.2 | leopard | volvo | machinegun |
|---|---|---|---|---|---|---|---|---|
| Linéaire | PE | 8.57 | 9.76 | 9.70 | 9.63 | 9.44 | 8.99 | 8.63 |
| | SD | 10.24 | 12.34 | 11.22 | 11.44 | 11.66 | 11.29 | 10.51 |
| MLP | PE | 6.89 | 7.88 | 8.55 | 8.07 | 7.92 | 7.27 | 7.13 |
| | SD | 10.09 | 11.60 | 10.72 | 11.00 | 10.78 | 10.96 | 10.36 |

Pour explorer les résultats, l'indépendance de la prédiction en fonction du type du bruit et du RSB, la PE a aussi été calculée pour chaque type de bruit et de RSB. De plus, afin d'évaluer l'indépendance de la prédiction en fonction du locuteur et du RSB, la PE a été calculée pour chaque locuteur et RSB.

Sur la figure 5, nous remarquons que la prédiction est liée au WER ciblé : les résultats sont plus précis lorsque le système obtient un WER faible ou important. Cette variabilité se constate car l'impact du modèle de langage sur le WER n'est pas ici quantifié. Par exemple, combien de phonèmes corrects sur un mot et ses voisins permettent d'identifier une suite de mots ? De plus, nous pouvons constater que le WER prédit est généralement sous estimé.

Nous remarquons aussi sur la figure 5 que le type de bruit n'influe pas sur la qualité de la prédiction : la méthode semble indépendante aux types de bruit.

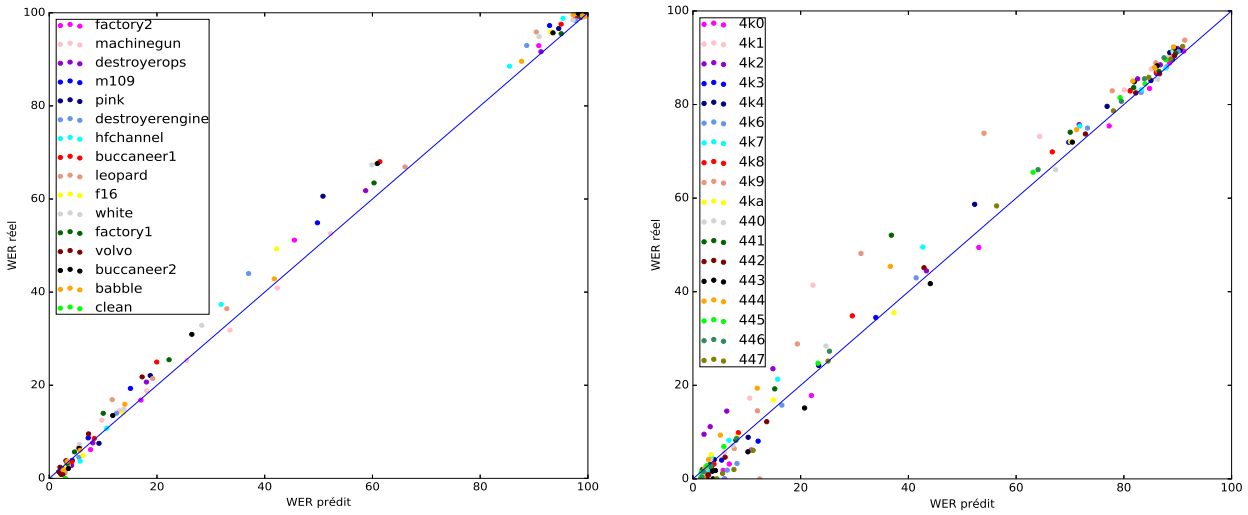



FIGURE 5 – À gauche, la différence entre la moyenne réelle et la prédiction (moyenne calculé pour le même type de bruit et de RSB). À droite, la différence entre la moyenne réelle et la prédiction (moyenne calculé pour le même locuteur et RSB).

Les locuteurs ont un impact sur les performances de la prédiction (voir figure 5). Par exemple, la prédiction est plus précise pour le locuteur 445 que pour le locuteur 4k9. Cependant, comme les textes dictés ne sont pas identiques entre les locuteurs, nous ne pouvons pas savoir si cet impact provient de la voix du locuteur, de la vitesse d'élocution ou du contenu des phrases lues.

# 5 Conclusions

Quel que soit la parole produite, le bruit va la dégrader. Connaître précisement l'impact du bruit sur la qualité de la transcription automatique de la parole, permet d'estimer un maximum ateignable par le système de RAP. Notre étude prouve qu'il est possible d'estimer *a priori* le WER obtenu par un système de RAP sur un fichier audio. En effet, les résultats montrent qu'avec une fenêtre temporelle suffisamment large, la prédiction est proche du WER réel (majoritairement inférieure à 8% d'erreur). Notre corpus étant composé de tours de parole indépendants, actuellement seul le calcul de la moyenne des WER prédits pour chaque tour de parole est proposé. Cependant, il est tout a fait possible d'imaginer une autre combinaison de scores. Par exemple, le filtrage des prédictions aberrantes ou l'utilisation d'un autre opérateur que la moyenne pour combiner les différents scores. Ce travail s'est focalisé uniquement sur les dégradations provoquées par le bruit afin d'analyser plus finement son impact sur le WER. Les paramètres extraits du bruit et de la parole séparée sont donc suffisamment corrélés au WER pour permettre une prédiction efficace pour de la parole lue.

La méthode de prédiction était, pour le moment, dépendante du locuteur afin d'analyser précisément l'impact du bruit. Une analyse de l'influence du locuteur sur le WER, comme la vitesse d'élocution ou le genre semble également intéressante. Cette seconde analyse permettrait d'obtenir idéalement une prédiction *a priori* indépendante du locuteur...

# Références


(1994). Nist speech quality assurance (spqa) package v2.3. [Online]. Available : https ://www.nist.gov/itl/iad/mig/tools.

(Philadelphia : Linguistic Data Consortium, 1994). CSR-II (WSJ1) complete LDC94S13A.

BROOKES M. (2006). Voicebox : A speech processing toolbox for matlab. [Online]. Available : https ://goo.gl/hVRjXZ.

EYBEN F. (2015). Real-time speech and music classification by large audio feature space extraction. p. 20–21. Springer.

GHANNAY S., ESTÈVE Y. & CAMELIN N. (2015). Word embeddings combination and neural networks for robustness in asr error detection. In *European Signal Processing Conference*.

HERMANSKY H., VARIANI E. & PEDDINTI V. (2013). Mean temporal distance : Predicting asr error from temporal properties of speech signal. In *Int. Conf. Acoust. Speech Signal Process* : IEEE.

JIANG H. (2005). Confidence measures for speech recognition : A survey. *Speech Communication*, p. 45 (4) 455–470.

JOHN, ET AL. G. (DVD. Philadelphia : Linguistic Data Consortium, 1993). CSR-I (WSJ0) complete LDC93S6A.

MEYER B., MALLIDI S., KAYSER H. & HERMANSKY H. (2017). Predicting error rates for unknown data in automatic speech recognition. In *Int. Conf. Acoust. Speech Signal Process* : IEEE.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P., SILOVSKY J., STEMMER G. & VESELY K. (2011). The kaldi speech recognitiion toolkit. In *Workshop on Automatic Speech Recognition and Understanding*, p. 1–4 : IEEE.

RUWEI L., YANAN L., YONGQIANG, L.AND LIANG L. & WEILI C. (2016). Ilmsaf based speech enhancement with dnn and noise classification. *Speech Communication*, **85**, 53–70.

VARGA A. & STEENEKEN H. (1993). Assessment for automatic speech recognition : II NOISEX-92 : A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, **12**, 247–251.

WANG D. (2005). On ideal binary mask as the computational goal of auditory scene analysis. *Speech Separation by Humans and Machines*, p. 181–197.

XU H., DALSGAARD P., TAN Z.-H. & LINDBERG B. (2007). Noise condition-dependent training based on noise classification and snr estimation. *IEEE transactions on audio, speech, and language processing*, **15**(8), 2431–2443.

YIMING S. & RUI W. (2015). Voice activity detection based on the improved dual-threshold method. In *Int. Con. on Intelligent Transportation, Big Data and Smart City*, p. 996–999.

ZWICKER E. (1961). Subdivision of the audible frequency range into critical bands. *The Journal of the Acoustical Society of America*, **33**, 248–248.

# Simulation d'erreurs de reconnaissance automatique dans un cadre de compréhension de la parole

Edwin Simonnet Sahar Ghannay Nathalie Camelin Yannick Estève
LIUM, Le Mans Université, France
firstname.lastname@univ-lemans.fr

RÉSUMÉ

Cet article propose une méthode de simulation d'erreurs de systèmes de reconnaissance automatique de la parole (SRAP) à partir de transcriptions manuelles, et montre son utilité pour rendre les systèmes de compréhension automatique de la parole (SCAP) plus robustes aux erreurs de SRAP. Partant du principe que le SRAP confond les mots acoustiquement et linguistiquement proches, cette méthode s'appuie sur l'utilisation de plongements de mots acoustiques et linguistiques pour calculer une mesure de similarité entre les mots : cette mesure vise à prédire les confusions de mots faites par le SRAP. Les expériences menées sur le corpus MEDIA (réservations d'hôtel) montrent que cette approche améliore significativement les performances des SCAP avec une réduction relative de 21,2% du taux d'erreur concept/valeur, en particulier quand le SCAP est neuronal (réduction de 22,4%). Une comparaison avec une méthode de bruitage naïf montre la pertinence de l'approche de bruitage proposée.

ABSTRACT

**Simulating ASR errors for training SLU systems**

This paper presents an approach to simulate automatic speech recognition (ASR) errors from manual transcriptions and how it can be used to improve the performance of spoken language understanding (SLU) systems. The proposed method is based on the use of both acoustic and linguistic word embeddings in order to define a similarity measure between words. This measure is dedicated to predict ASR confusions. Actually, we assume that words acoustically and linguistically close are the ones confused by an ASR system. Experiments were carried on the French MEDIA corpus focusing on hotel reservation. They show that this approach significantly improves SLU system performance with a relative reduction of 21.2% of concept/value error rate (CVER), particularly when the SLU system is based on a neural approach (reduction of 22.4% of CVER). A comparison to a naive noising approach shows that the proposed noising approach is particularly relevant.

MOTS-CLÉS : compréhension de la parole, augmentation des données, bruitage, reconnaissance automatique de la parole, erreurs.

KEYWORDS: spoken language understanding, data augmentation, noising, automatic speech recognition, errors.

## 1 Introduction

Les systèmes de compréhension de la parole (SCAP) ont pour but l'extraction d'informations sémantiques dans un discours. Dans un système de dialogue, cela consiste à extraire automatiquement des concepts sémantiques sous forme de couples concept/valeur à partir de transcriptions automatiques afin d'alimenter le gestionnaire de dialogue.Nous considérons ainsi la tâche de compréhension de


10.21437/JEP.2018-30

la parole comme une tâche de traduction où les séquences d'hypothèses de mots issues d'un SRAP doivent être traduites en une séquence de concepts sémantiques associés à leurs valeurs. Ainsi, la bonne performance du système de compréhension est donc fortement liée à la bonne performance du système de transcription. En effet, les erreurs de transcription sont susceptibles d'affecter les mots supports d'un concept, rendant difficile à la fois la détection du concept et l'extraction de sa valeur.

Dans l'optique de rendre plus robustes les systèmes de compréhension aux erreurs de reconnaissance, il est habituel d'entrainer le modèle sur des transcriptions automatiques plutôt que sur des transcriptions manuelles. Comme les corpus requis pour l'apprentissage des systèmes de dialogue sont rares, certaines méthodes ont été proposées pour simuler les erreurs de transcription dans ce cadre (Pietquin & Beaufort, 2005; Schatzmann *et al.*, 2007). La simulation d'erreurs de transcriptions a également été utilisée pour l'entraînement de modèles de langage discriminatifs afin d'améliorer les performances des SRAP en terme de taux d'erreurs mots (Jyothi & Fosler-Lussier, 2010).

De nos jours, les SCAP sont souvent construits avec une approche guidée par les données (Sarikaya *et al.*, 2014; Mesnil *et al.*, 2015; Hakkani-Tür *et al.*, 2016). Des annotations manuelles sont habituellement produites pour étiqueter des transcriptions manuelles avec des étiquettes sémantiques afin de construire un corpus d'apprentissage. Dans l'étude présentée ici nous supposons – et le vérifions – que la construction des SCAP à partir de transcriptions automatiques est une bonne solution pour les rendre plus robustes aux erreurs de transcriptions. Or, l'obtention des transcriptions automatiques nécessite d'avoir à disposition d'une part des enregistrements audio relatifs aux annotations sémantiques et d'autre part un SRAP. Afin que ce dernier soit efficace, il nécessite lui aussi des données d'apprentissage et de validation, ces dernières étant souvent les mêmes que celles utilisées pour l'apprentissage et la validation SCAP. Il convient donc de manipuler ces données avec prudence afin d'éviter des biais et notamment celui du sur-apprentissage.

Dans le cadre de la construction d'un SCAP performant, cette étude propose une approche de simulation des erreurs de reconnaissance à partir des transcriptions manuelles afin d'une part de s'affranchir de la nécessité de données audio et d'un SRAP lors de la phase d'apprentissage et d'autre part d'avoir néanmoins à disposition un corpus proche de celui à gérer lors du déploiement. Notre approche consiste à simuler et introduire des erreurs dans les transcriptions manuelles en substituant des mots corrects par des mots similaires. Nous supposons que les mots susceptibles d'être confondus par un SRAP sont des mots acoustiquement proches. Cette hypothèse a également été retenue dans (Fosler-Lussier *et al.*, 2002; Stuttle *et al.*, 2004), où la simulation des erreurs est basée sur la similarité phonétique des mots pour évaluer leur similarité. De plus, nous considérons que ces mots confondus sont également linguistiquement proches.

Pour calculer une mesure de similarité entre les mots, nous présentons une nouvelle approche utilisant des plongements de mots acoustiques et linguistiques. Dans nos expériences, nous évaluons l'impact de cette approche en bruitant le corpus d'apprentissage de deux SCAP : un basé sur des champs aléatoires conditionnels (Lafferty *et al.*, 2001) (CRF) et l'autre sur un réseau de neurone récurent bidirectionnel encodeur-décodeur avec un mécanisme d'attention (Cho *et al.*, 2014) (RNN-EDA). Ces expériences sont menées sur le corpus français MEDIA, sur lequel les CRF fonctionnent toujours mieux que les approches neuronales (Vukotic *et al.*, 2015; Simonnet *et al.*, 2017).

# 2 Mesure de similarité et simulation d'erreurs de SRAP

Nous proposons une mesure de similarité qui s'appuie sur l'utilisation des plongements linguistiques et acoustiques pour prédire une liste de mots qui pourraient être substitués par un système de reconnaissance de la parole à un mot effectivement prononcé. Nous nommons cette liste une *liste de*

*confusion*. Elle se compose des mots les plus proches du mot analysé selon une mesure de similarité qui s'appuie sur la combinaison des similarités cosinus des plongements de types linguistique et acoustique.

Les plongements linguistiques de mots correspondent à la combinaison par analyse en composante principale de différents types de plongement de mots : *word2vecf* (Levy & Goldberg, 2014), *skip-gram* fournis par *word2vec* (Mikolov *et al.*, 2013), et *GloVe* (Pennington *et al.*, 2014), comme décrit dans (Ghannay *et al.*, 2016).

Les plongement acoustiques de mots correspondent à la projection de séquences acoustiques de longueur variable dans un espace de faible dimension de telle sorte que les mots qui se prononcent de la même manière sont projetés dans la même zone, tandis que les mots qui se prononcent différemment sont projetés dans des zones différentes. L'approche que nous avons utilisée pour construire ces représentations s'inspire de celle proposée dans (Bengio & Heigold, 2014).

## 2.1 Interpolation linéaire des similarités linguistique et acoustique

Dans cette étude, nous proposons d'utiliser des plongements linguistiques et acoustiques pour prédire les confusions faites par le SRAP. Pour construire une mesure de similarité combinant des plongements de mots de natures différentes, nous proposons d'utiliser l'interpolation linéaire des similarités cosinus linguistique et acoustique. La similarité résultante est appelée $LA_{SimInter}$, et est définie comme suit :

$$LA_{SimInter}(\lambda, w_1, w_2) = (1 - \lambda) \times L_{Sim}(w_1, w_2) + \lambda \times A_{Sim}(w_1, w_2) \quad (1)$$

où $w_1$ et $w_2$ sont les deux mots à comparer et $\lambda$ est le coefficient d'interpolation. Les similarités $L_{Sim}$ et $A_{Sim}$ sont calculées avec la similarité cosinus appliquée respectivement aux plongements linguistiques et acoustiques de $w_1$ et $w_2$.

Comme notre objectif est de prédire ou corriger les erreurs du SRAP, nous voulons optimiser la valeur $\lambda$ à cette fin. Pour estimer $\lambda$, une liste connue d'erreurs de substitution générées par le SRAP est utilisée. Dans cette liste, nous définissons $h$ comme étant l'hypothèse de mot erronée et $\overline{r}$ le mot de référence qui a été substitué par $h$. Pour chaque paire de mots $(h, \overline{r})$ dans la liste, nous calculons la probabilité que le mot $h$ soit reconnu lorsque le mot de référence $\overline{r}$ est erroné : $P(h|\overline{r}) = \frac{\#(h,\overline{r})}{\#\overline{r}}$, où $\#(\mathrm{h}, \overline{r})$ est le nombre de substitutions de $\overline{r}$ par $h$ et $\#\overline{r}$ le nombre d'erreurs sur le mot de référence $\overline{r}$.

Nous proposons alors de retenir le coefficient d'interpolation $\hat{\lambda}$ qui minimise l'erreur quadratique moyenne (MSE) entre la valeur proposée par $LA_{SimInter}(\lambda, h, \overline{r})$ et la valeur effective de $P(h|\,\overline{r})$. Nous définissons alors $\hat{\lambda}$ tel que :

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}}\, MSE(\forall(h, \overline{r}) : P(h|\overline{r}), LA_{SimInter}(\lambda, h, \overline{r})) \quad (2)$$

où $LA_{SimInter}(\lambda, h, \overline{r})$ et $P(h|\,\overline{r})$ sont calculés sur tous les couples ($h$, $\overline{r}$) possibles.

En utilisant $LA_{SimInter}$ avec $\hat{\lambda}$, il est maintenant possible de proposer pour un mot donné sa liste de confusion contenant ses voisins les plus proches linguistiquement et acoustiquement. La valeur de $LA_{SimInter}(\hat{\lambda}, x, y)$ est considérée comme une mesure de similarité entre les mots $x$ et $y$ et nous la notons plus simplement $confus(x, y)$.

## 2.2 La simulation d'erreurs

Pour simuler des erreurs de reconnaissance, on applique la mesure de similarité $confus(x, y)$ afin de substituer dans la transcription manuelle des mots corrects par des mots erronés de la liste de

confusion.

En déterminant un taux d'erreur $e$ (uniquement composés de substitutions), on modifie aléatoirement un pourcentage $e$ des occurrences de mot. Ces substitutions sont faites après avoir défini deux seuils : le seuil $c$ qui réfère la valeur la plus basse de $confus(\overline{r}, h)$ qui permet de substituer le mot $\overline{r}$ par le mot $h$, et le seuil $n$ qui limite le nombre de substitutions possibles de $\overline{r}$ parmi les $n$ mots $h_i$ les plus proches (*i.e.* les mots $h_i$ tels que la valeur $confus(\overline{r}, h_i)$ est l'une des $n$ valeurs les plus hautes étant donné $\overline{r}$). Le mot $h$ est choisi aléatoirement dans la liste des mots $h_i$ qui respectent les contraintes des seuils $n$ et $c$.

# 3 Protocole Expérimental

Cette partie décrit le protocole expérimental inspiré d'une étude précédente (Simonnet *et al.*, 2017).

## 3.1 Le corpus MEDIA

Le corpus utilisé est le corpus MEDIA, collecté dans le projet français Media/Evalda (Bonneau-Maynard *et al.*, 2005). Il contient trois ensembles de dialogues téléphoniques humain/ordinateur liés au tourisme, à savoir : un ensemble d'apprentissage (APP) avec environ 17,7k phrases, un ensemble de développement (DEV) avec 1,3k phrases et un ensemble d'évaluation (TEST) contenant 3,5k phrases. Le corpus a été annoté manuellement avec des concepts sémantiques caractérisés par une étiquette et sa valeur. Les évaluations sont effectuées avec les ensembles DEV et TEST et rapportent les taux d'erreur CER (concept error rate) pour les étiquettes de concepts seulement et les taux d'erreur CVER (concept-value error rate) pour les paires étiquette-valeur. Il est à noter que le nombre de concepts annotés dans une phrase a une grande variabilité et peut inclure plus de 30 concepts annotés.

Pour ces expériences, une variante du SRAP développé par le LIUM est utilisée. Elle a remporté la dernière campagne d'évaluation sur la langue française (Rousseau *et al.*, 2014). Ce système est basé sur la boîte à outils de reconnaissance vocale Kaldi (Povey *et al.*, 2011). Une description détaillée du SRAP est donnée dans (Simonnet *et al.*, 2017). Les taux d'erreur mot pour les corpus APP, DEV et TEST sont respectivement de 23,7%, 23,4% et 23,6%.

## 3.2 Descriptions des systèmes de compréhension

Deux systèmes de compréhension sont comparés sur le corpus MEDIA. Le premier est un RNR-EDA similaire à celui utilisé pour la traduction automatique proposé dans (Cho *et al.*, 2014). Le second est basé sur des CRF. Les deux architectures construisent leur modèle d'apprentissage sur le même ensemble de descripteurs en entrée, avec des valeurs continues pour le premier et des valeurs discrètes pour le second.

### 3.2.1 Descripteurs de mot

Afin d'améliorer les performances de compréhension des systèmes, un ensemble de descripteurs, inspiré de (Hahn *et al.*, 2011), représente chaque occurence de mot en entrée des SCAP. Il s'agit de : le mot, la catégorie sémantique prédéfinie qui peut être spécifique à MEDIA ou plus générale ; des caractéristiques syntaxiques et morphologiques ; et deux mesures de confiance : la probabilité *postérieure* (*pap*) et la mesure de confiance issue d'un perceptron multi-couche. Ces deux dernières caractéristiques estiment la fiabilité du mot reconnu par le SRAP. La description détaillée de ces descripteurs se trouve dans (Simonnet *et al.*, 2017).

Les deux SCAP prennent en entrée tous ces descripteurs à l'exception des mesures de confiance où seulement une est gardée dans un but de cohérence expérimentale comme cela sera décrit dans la sous-section 3.3. Ces architectures doivent également être calibrées sur leurs hyper-paramètres respectifs afin de donner les meilleurs résultats. La façon dont la meilleure configuration est choisie est décrite dans la section 4.

#### 3.2.2 Système de compréhension de la parole basé sur les RNR-EDA

Le RNR-EDA proposé, inspiré d'une architecture de traduction automatique, a été implémenté à partir de l'outil *nmtpy* (Caglayan *et al.*, 2017). L'étiquetage de concept est considéré comme une traduction de mots (langage source) vers étiquettes sémantiques (langage cible). Une description détaillée du RNR-EDA est donnée dans (Simonnet *et al.*, 2017).

#### 3.2.3 Système de compréhension de la parole basé sur les CRF

Les expériences passées décrites dans (Hahn *et al.*, 2011) ont montré que les meilleures performances en annotation sémantique sur les transcriptions manuelles et automatiques du corpus MEDIA ont été obtenues avec les CRF. Plus récemment, dans (Vukotic *et al.*, 2015), cette architecture a été comparée à un RNR bidirectionnel (biRNR). La conclusion fut que les CRF surpassent les biRNR sur le corpus MEDIA, alors que de meilleurs résultats ont été observés par les biRNR sur le corpus ATIS (Hemphill *et al.*, 1990). Ceci s'explique probablement par le fait que MEDIA contient des contenus sémantiques dont les mentions sont plus difficiles à désambiguïser, et les CRF exploitent plus efficacement des contextes complexes ((Vukotic *et al.*, 2015)).

Par soucis de comparaison avec le meilleur SCAP proposé dans (Hahn *et al.*, 2011), la boîte à outils Wapiti (Lavergne *et al.*, 2010) a été utilisée dans notre étude. Néanmoins, l'ensemble des descripteurs utilisés par le système proposé dans cet article est différent de celui utilisé dans (Hahn *et al.*, 2011). Parmi les nouveautés utilisées dans notre système, nous considérons des descripteurs syntaxiques et des mesures de confiance de SRAP et notre modèle de configuration est différent. Après de nombreuses expériences effectuées sur le DEV, notre modèle de descripteur final inclut les instances précédentes et suivantes pour les mots et la catégorie grammaticale dans un unigram ou un bigram afin d'associer une étiquette sémantique avec le mot en cours. De plus sont associés avec le mot courant les catégories sémantiques des deux instances précédentes et des deux suivantes. Les autres descripteurs ne sont considérés qu'à la position courante. De plus, l'outil discretize4CRF [1] est utilisé pour discrétiser les mesures de confiance de SRAP afin qu'elles soient acceptées en entrée des CRF.

### 3.3 Simulation d'erreurs de transcriptions

La méthode présentée dans la sous-section 2 est appliquée afin de simuler des erreurs de SRAP. À partir des annotations manuelles du corpus MEDIA, nous construisons différents ensembles de données. Dans ces simulations, nous avons fixé la valeur de $e$ à 20%, ce qui représente le taux de mots que nous corrompons au hasard dans les transcriptions manuelles.

Deux simulations différentes ont été testées, en choisissant différentes valeurs de seuil $n$ et $c$ ;

— **corpus B.7** : $n = 7$ et $c = 0.4$ ;
— **corpus B.10** : $n = 10$ et $c = 0.5$.

Un autre ensemble de données artificiel a été créé, appelé **corpus B.n** : ce corpus ne prend pas en compte la mesure de similarité. Dans cet ensemble de données, le même pourcentage de mots $e = 20\%$

1. https ://gforge.inria.fr/projects/discretize4crf/

issus des transcriptions manuelles est substitué de manière aléatoire, en choisissant simplement un mot au hasard dans l'ensemble du vocabulaire MEDIA. Quand un mot correct est remplacé par un mot confondu, nous utilisons la mesure de similarité comme mesure de confiance de SRAP.

Dans un but de cohérence expérimentale, lorsque nous travaillons sur des sorties de SRAP, nous donnons seulement une mesure de confiance parmi les deux disponibles afin d'avoir toujours le même nombre de mesures de confiance dans tous les cas.

# 4 Résultats Expérimentaux

Pour les deux SCAP, l'apprentissage est fait sur l'APP et les meilleures configurations sont choisies pour optimiser le CVER sur le DEV. Les résultats sur le TEST en CER et CVER sont reportés dans les tables 1 et 2, où **M** fait référence au corpus manuel, **A** à un corpus composé de transcriptions automatiques, et **B** à un corpus bruité. Le TEST est constitué uniquement de transcriptions automatiques, alors que la nature des corpus APP ou DEV varie dans nos expériences.

## 4.1 Analyse de l'apport des transcriptions bruitées à l'apprentissage

Puisque l'évaluation sur le TEST est faite sur des transcriptions automatiques, nous considérons dans un premier temps qu'un corpus DEV composé de transcriptions automatiques est également disponible. Ce corpus est moins difficile à collecter qu'un corpus d'entraînement (1.3k phrases *vs.* 17.7k) et sa manipulation n'entraine ni biais, ni sur-apprentissage. Les résultats expérimentaux de cette configuration sont visibles dans la table 1.

| | *APP* | M | A | B.7 | B.7 x2 | M +B.7 | M +B.10 | M +B.n | M +A | M +B.7+A |
|---|---|---|---|---|---|---|---|---|---|---|
| | *DEV* | A | A | A | A | A | A | A | A | A |
| *RNR EDA* | *CER* | 31,6 | 22,5 | 23,8 | 23,2 | 22,7 | 23,3 | 23,7 | 20,7 | 20,2 |
| | *CVER* | 36,2 | 28,3 | 29 | 28,8 | 28,1 | 28,5 | 28,8 | 25,8 | 26 |
| *CRF* | *CER* | 27,5 | 19,9 | 22,6 | 26,3 | 22,6 | 23,2 | 25 | 20,2 | 29,1 |
| | *CVER* | 31,6 | 25,1 | 27,7 | 31,3 | 27,7 | 28,3 | 30,3 | 25,3 | 33 |

TABLE 1 – Comparaison de différents APP en CER et CVER sur un TEST et un DEV automatique.

Nous pouvons d'abord noter que notre hypothèse sur l'importance d'apprendre sur des données proches des données de test (avec des transcriptions automatiques ou contenant des simulations d'erreurs) est vérifiée : avec l'APP **A**, les résultats des RNR-EDA et des CRF sont significativement meilleurs que ceux fait avec un APP **M**. On voit également que les CRF surpassent significativement les RNR-EDA sus les corpus d'entraînement **M** et **A**. Il est également clair que l'entraînement d'un SCAP sur des transcriptions manuelles est largement insuffisant pour gérer les transcriptions automatiques. Le système doit être préparé aux erreurs de transcriptions.

L'entraînement sur un corpus bruité (colonne B.7) obtient des résultats intéressants. On obtient une nette amélioration par rapport aux mauvais résultats obtenus sur les transcriptions manuelles seulement. Il se rapproche des résultats utilisant les transcriptions automatiques pures et confirme ainsi que notre approche pour simuler des erreurs de transcription est adaptée à cette tâche. Entraîner sur un corpus bruité doublé (colonne double B.7, dans laquelle deux simulations d'erreurs de SRAP successives sur l'APP ont été utilisées) permet d'améliorer un peu les résultats sur le RNR-EDA tout en aggravant fortement ceux des CRF.

De meilleurs résultats peuvent être obtenus en combinant des corpus manuels et bruités. En utilisant l'ensemble de données B.7 combiné au manuel, les résultats sont tout aussi bons que des transcriptions automatiques pures pour le RNR-EDA. Les CRF obtiennent les mêmes résultats que pour B.7 seulement.

Nous pouvons également comparer les différents types de bruit. Le B.7 obtient de meilleurs résultats que le B.10, ce qui montre qu'en substituant des mots corrects à des mots globalement moins semblables, les résultats diminuent. De plus, même si l'application de bruit naïf (B.n) obtient de meilleurs résultats que l'utilisation de transcriptions manuelles (APP M), nous obtenons les plus mauvais scores parmi les approches bruitées. Ceci montre l'importance d'un bruit généré intelligemment, et valide implicitement notre approche de simulation d'erreurs de transcription.

Finalement, les meilleurs résultats obtenus qui surpassent les transcriptions automatiques pures (A) sont obtenus en entraînant les SCAP sur une combinaison de sorties automatiques et manuelles (M+A). Les deux SCAP trouvent leur meilleure performance dans cette configuration et l'écart entre CRF et RNR-EDA a été fortement réduit par rapport aux expériences sur A ou M seulement. L'entraînement sur une triple combinaison de corpus manuel, automatique et bruité n'augmente pas davantage ces résultats.

En général, les CRF surpassent significativement les RNR-EDA lorsque ces systèmes sont entraînés sur un corpus manuel ou automatique. Mais les RNR-EDA tirent meilleur parti de la simulation d'erreurs, ou de la combinaison manuelle et automatique par rapport aux CRF. Au final, les meilleurs résultats des RNR-EDA et CRF sont très proches, montrant un potentiel des réseaux de neurones, non partagé par les CRF, à apprendre des informations pertinentes à partir de données bruitées.

## 4.2 Apprentissage sans transcriptions automatiques

Dans cette section, nous explorons le scénario dans lequel aucune donnée issue d'un SRAP n'est disponible pour entraîner le système de compréhension (DEV inclus). Cela peut devenir problématique lorsque le système de compréhension doit effectuer des phases de validation durant le processus d'apprentissage, ce qui est le cas des RNR-EDA. Les CRF pour leur part n'utilisent pas le DEV pendant l'entraînement (la configuration optimale n'est pas modifiée et les scores des CRF restent inchangés). Ainsi, les résultats visibles dans la table 2 ne concernent que les RNR-EDA.

| | *APP* | M | B.7 | M+B.7 |
|---|---|---|---|---|
| | *DEV* | M | B.7 | B.7 |
| *RNR EDA* | *CER* | 33,9 | 23,5 | 23,1 |
| | *CVER* | 38,2 | 28,6 | 28,5 |

TABLE 2 – Comparaison en CER et CVER obtenus sur un TEST automatique mais sans données automatiques pour l'APP ou le DEV.

En général, sauf pour l'APP bruité seul, de meilleurs résultats sont atteints en validant sur un DEV automatique, plus proche des données de TEST. Néanmoins, même si ces résultats sont un peu moins bons que ceux obtenus en validant sur un DEV automatique, on peut remarquer qu'il est possible d'améliorer très significativement les performances des SCAP en appliquant notre approche de simulation d'erreurs pour enrichir ou bruiter les données d'apprentissage et de développement des SCAP ne disposant que de transcriptions manuelles.

# 5 Conclusion

Deux architectures de compréhension de la parole basées sur des RNR-EDA et des CRF ont été comparées dans cette étude. Une simulation d'erreur de transcription basée sur une mesure de similarité construite à partir de plongements de mots acoustiques et linguistiques a été proposée et utilisée pour bruiter un corpus manuel annoté. Les expériences montrent que ce bruitage est pertinent pour enrichir et préparer un corpus d'entraînement de SCAP. Si aucun SRAP n'est disponible pour préparer ces données, notre proposition offre une amélioration très significative des performances des SCAP, de 36,2% de CVER avec seulement des annotations manuelles dans le corpus d'entraînement, contre 28,5% de CVER en appliquant notre approche : ceci représente une réduction relative de 21,2% des erreurs en concept-valeur. Un autre résultat intéressant dans cette étude est la diminution des écarts, en terme de CER ou de CVER, entre CRF et RNR-EDA sur le corpus MEDIA. Aucun changement n'a été fait sur ce corpus depuis 2011 (Hahn *et al.*, 2011) et les CRF sont toujours dominants. Nos résultats montrent qu'il est maintenant possible d'obtenir des résultats similaires avec une architecture neuronale. Nous nous attendons à proposer de nouvelles contributions pour rendre les réseaux de neurones plus efficaces que les CRF, qui ont atteint un plateau il y a plusieurs années sur cette tâche. Dans un avenir proche, nous considérerons également d'autres approches de simulation d'erreurs de SRAP pour comparer leur impact aux nôtres afin de préparer et d'enrichir le corpus d'entraînement des SCAP. Nous expérimenterons également l'utilisation de notre simulation de SRAP sur d'autres tâches, comme la détection d'erreurs de SRAP par exemple.

# Références


BENGIO S. & HEIGOLD G. (2014). Word embeddings for speech recognition. In *INTERSPEECH*, p. 1053–1057.

BONNEAU-MAYNARD H., ROSSET S., AYACHE C., KUHN A. & MOSTEFA D. (2005). Semantic annotation of the french media dialog corpus. In *Ninth European Conference on Speech Communication and Technology*.

CAGLAYAN O., GARCÍA-MARTÍNEZ M., BARDET A., ARANSA W., BOUGARES F. & BARRAULT L. (2017). Nmtpy : A flexible toolkit for advanced neural machine translation systems. *arXiv preprint arXiv :1706.00457*.

CHO K., VAN MERRIENBOER B., GULCEHRE C., BOUGARES F., SCHWENK H. & BENGIO Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*.

FOSLER-LUSSIER E., AMDAL I. & KUO H.-K. J. (2002). On the road to improved lexical confusability metrics. In *ISCA Tutorial and Research Workshop (ITRW) on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*.

GHANNAY S., FAVRE B., ESTEVE Y. & CAMELIN N. (2016). Word embedding evaluation and combination. In *of the Language Resources and Evaluation Conference (LREC 2016), Portoroz (Slovenia)*, p. 23–28.

HAHN S., DINARELLI M., RAYMOND C., LEFEVRE F., LEHNEN P., DE MORI R., MOSCHITTI A., NEY H. & RICCARDI G. (2011). Comparing stochastic approaches to spoken language understanding in multiple languages. *IEEE Transactions on Audio, Speech, and Language Processing*, **19**(6), 1569–1583.

HAKKANI-TÜR D., TUR G., CELIKYILMAZ A., CHEN Y.-N., GAO J., DENG L. & WANG Y.-Y. (2016). Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In *Proceedings of The 17th Annual Meeting of the International Speech Communication Association*.

HEMPHILL C. T., GODFREY J. J., DODDINGTON G. R. *et al.* (1990). The atis spoken language systems pilot corpus. In *Proceedings of the DARPA speech and natural language workshop*, p. 96–101.

JYOTHI P. & FOSLER-LUSSIER E. (2010). Discriminative language modeling using simulated asr errors. In *Eleventh Annual Conference of the International Speech Communication Association*.

LAFFERTY J., MCCALLUM A., PEREIRA F. *et al.* (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, p. 282–289.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 504–513 : Association for Computational Linguistics.

LEVY O. & GOLDBERG Y. (2014). Dependency based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, p. 302–308.

MESNIL G., DAUPHIN Y., YAO K., BENGIO Y., DENG L., HAKKANI-TUR D., HE X., HECK L., TUR G., YU D. *et al.* (2015). Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, **23**(3), 530–539.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. In *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, volume 12.

PIETQUIN O. & BEAUFORT R. (2005). Comparing asr modeling methods for spoken dialogue simulation and optimal strategy learning. In *Ninth European Conference on Speech Communication and Technology*.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P. *et al.* (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number EPFL-CONF-192584 : IEEE Signal Processing Society.

ROUSSEAU A., BOULIANNE G., DELÉGLISE P., ESTÈVE Y., GUPTA V. & MEIGNIER S. (2014). LIUM and CRIM ASR system combination for the REPERE evaluation campaign. In *International Conference on Text, Speech, and Dialogue*, p. 441–448 : Springer.

SARIKAYA R., HINTON G. E. & DEORAS A. (2014). Application of deep belief networks for natural language understanding. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, **22**(4), 778–784.

SCHATZMANN J., THOMSON B. & YOUNG S. (2007). Error simulation for training statistical dialogue systems. In *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, p. 526–531 : IEEE.

SIMONNET E., GHANNAY S., CAMELIN N., ESTEVE Y. & RENATO D. M. (2017). ASR error management for improving spoken language understanding. In *INTERSPEECH*.

STUTTLE M., WILLIAMS J. & YOUNG S. (2004). A framework for dialog systems data collection using a simulated asr channel. In *ICSLP 2004*.

VUKOTIC V., RAYMOND C. & GRAVIER G. (2015). Is it time to switch to word embedding and recurrent neural networks for spoken language understanding ? In *InterSpeech*.

XXXIIe Journées d'Études sur la Parole
4-8 juin 2018, Aix-en-Provence, France

# Simulation numérique des apériodicités vocales dues aux fluctuations de la tension musculaire

Jean Schoentgen[1, 2] Dhouha Rezgui[2] Francis Grenez[2]
(1) Fonds de la Recherche Scientifique, Rue d'Egmont 5, B - 1000 Bruxelles, Belgique
(2) B.E.A.M.S., Université Libre de Bruxelles, CP165/51, 50, Av. F.-D. Roosevelt, B-1050 Bruxelles, Belgique
jschoent@ulb.ac.be, rezgui_dhouha@yahoo.com,fgrenez@ulb.ac.be

RÉSUMÉ

L'objectif est le dévelopement et l'étude d'un modèle du jitter vocal destiné à la sythèse numérique de la qualité de voix. Le modèle est numériquement stable et compact et évite les hypothèses *ad hoc* d'un modèle existant. Le jitter vocal est obtenu à partir des fluctuations de la tension musculaire causées par les contractions élémentaires superposées du muscle thyro-aryténoïdien impliquant l'activité concomitante de plusieurs unités motrices. Les paramètres de contrôle sont le nombre d'unités motrices, le temps mort et le taux d'émission des neurones moteurs, ainsi que le temps latent et le temps de montée des contractions musculaires élémentaires. La présentation inclut une comparaison avec un modèle existant ainsi qu'une étude de l'influence des paramètres physiologiques sur un indice connu du jitter de la fréquence vocale.

ABSTRACT

**Numerical simulation of vocal aperiodicities owing to muscle tension fluctuations.**

The presentation is devoted to a model of vocal jitter, which is inspired by physiology. The simulation is numerically stable and compact so that it may be used in speech synthesis. Synthetic vocal jitter is obtained via simulated muscle tension fluctuations, which are the outcome of the superposition of the TA muscle twitches that involve the activity of several motor units. The control parameters are the number of active motor units, the dead time and firing rate of the motor neurones, as well as the rise time and latency of the muscle twitches. The presentation includes a comparison with an existing model as well as an analysis of the dependence of a popular cue of vocal jitter on physiological parameters.

MOTS-CLÉS : Jitter vocal, fluctuations de la tension musculaire, unités motrices, neurones moteurs.

KEYWORDS: Vocal frequency jitter, muscle tension fluctuations, motor units, motor neurons.

# 1 Introduction

L'objectif est la présentation d'un modèle des perturbations de la fréquence vocale qui est inspiré de la physiologie du muscle thyro-arythénoïdien, la comparaison à un modèle existant (Titze, 1991) et l'étude de l'influence des paramètres physiologiques sur des indices connus des perturbations de la fréquence vocale $Fo$. Le modèle est numériquement compact et exempt de régimes physiquement ou physiologiquement impossibles. Le temps de calcul permet son utilisation pour la synthèse vocale.


10.21437/JEP.2018-31

Kreiman a publié une énumération des causes possibles du jitter vocal, dans le sens large du terme, qui renvoie à une vaste gamme de vibrations irrégulières dont la majorité est mieux connue sous d'autres noms (Kreiman & Sidtis, 2011).
Le modèle qui est discuté ici simule les causes du jitter et du scintillement de la fréquence vocale lorsque les vrais plis vocaux vibrent pseudo-périodiquement et monodiquement dans un même mécanisme phonatoire. Lors de l'émission d'une voyelle soutenue, les perturbations de la périodicité stricte, des plus lentes aux plus rapides, sont les suivantes.

— La dérive ou la déclinaison de $Fo$.
— Le pleurage vocal qui désigne le tremblement physiologique causé par le flux sanguin pulsé et la respiration.
— Le tremblement vocal d'origine neurologique qui est dû à des taux d'émission non stationnaires des potentiels d'action des neurones moteurs.
— Le scintillement et le jitter vocal qui sont dus à la tension fluctuante des muscles laryngés.

R. Cook écrit que les régions dans le spectre du contour intonatif en-dessous et au-dessus du pic de vibrato sont appelées les régions du pleurage et scintillement vocal respectivement (Cook, 1999). Titze (1994) précise que le pleurage et le scintillement sont des modulations dans l'intervalle de 1 à 2 $Hz$ et de 10 à 12 $Hz$ respectivement. Le scintillement vocal n'est que rarement discuté dans la littérature, mais on sait qu'il influence les indices de perturbations courants (Alzamendi & Schlotthauer, 2017). Il apparaît ensemble avec le jitter dans des simulations inspirées de la physiologie du muscle vocal, ce qui suggère qu'ils ont une origine commune. En pratique, le jitter vocal rapporte les perturbations à court terme tandis que le scintillement vocal observé s'étale sur quelques cycles.
Nous utilisons ici le concept de jitter vocal dans l'acception courante du terme qui désigne la variabilité de cycle-à-cycle de la fréquence vocale ou des durées des cycles glottiques. Les indices acoustiques qui rapportent le jitter de la fréquence vocale sont à la fois populaires et critiqués. Ils sont populaires car ils ont la réputation d'être des marqueurs distinctifs de la qualité vocale d'une majorité de locuteurs. Ils sont critiqués car les utilisateurs perdent parfois de vue les conditions sous lesquelles les indices courants peuvent être obtenus fiablement, mais aussi parce que ces conditions suggèrent que le jitter vocal, dans le sens restreint utilisé ici, contribue de façon négligeable aux qualités plus extrêmes ou spectaculaires.
Les fluctuations de la tension du muscle thyro-arythénoïdien (TA) sont à même de causer des perturbations audibles de la fréquence vocale lorsque les vrais plis vocaux vibrent exclusivement, pseudo-périodiquement et monodiquement. On s'attend à ce que les fluctuations de la tension d'autres muscles, le crico-thyroïdien (CT) par exemple, contribuent moins car l'inertie du cartilage thyroïde est susceptible de les lisser.
La tension d'un muscle squelettique est l'effet de l'activité simultanée de plusieurs unités motrices. Une unité motrice est composée d'un neurone moteur qui innerve un groupe de fibres musculaires qui se contractent ensemble à l'arrivée d'une impulsion électrique, aussi appelée potentiel d'action, émise par le neurone. Cette contraction simultanée de plusieurs fibres est appelée une contraction élémentaire (ou secousse). La tension musculaire totale est la résultante de la co-existence de nombreuses secousses qui se chevauchent dans l'espace (suite à l'activité simultanée de plusieurs unités motrices) et dans le temps (suite à la succession rapide des potentiels d'action d'un seul neurone moteur.)

# 2 Le modèle "$\mathcal{N} - P$" des sources neurologiques des apériodicités vocales

Le modèle est ici appelé "$\mathcal{N} - P$" parce qu'il repose sur une perturbation aléatoire normale "$\mathcal{N}$" des positions "$P$" dans le temps des contractions musculaires élémentaires. En effet, Titze à proposé en 1991 un modèle des fluctuations de la tension du muscle TA. Les composants en sont (i) un modèle de la contraction élémentaire, (ii) la superposition des contractions élémentaires qui se suivent dans le temps simulant l'activité d'une seule unité motrice et (iii) la superposition asynchrone de plusieurs séquences de contractions simulant l'activité concomitante de plusieurs unités motrices.
Le modèle de la contraction élémentaire comprend un paramètre qui est le temps de monté $T_m$ (Herman, 2007; Titze, 1991). La contraction $s(t)$ proprement dite démarre à l'instant $t = 0$ et son amplitude est égale à l'unité.

$$s(t) = \frac{t}{T_m} e^{1 - \frac{t}{T_m}} \tag{1}$$

Les positions $t_j$ des contractions dues aux potentiels d'action d'un seul neurone moteur sont obtenues en centrant une distribution normale $\mathcal{N}$ sur des positions équidistantes d'une durée $\mu$ qui est égale à l'inverse du taux d'émission $\lambda$ en $Hz$ d'un neurone moteur. L'écart-type $\sigma$ des perturbations est fixée à l'aide du coefficient de variation $\nu = \sigma/\mu$ de la durée moyenne $\mu$ entre potentiels d'actions. La durée $T_l$ est le temps de latence qui sépare l'arrivée du potentiel d'action et le début de la contraction élémentaire.

$$t_j = \mathcal{N}(j \times \mu, \nu \times \mu) + T_l \tag{2}$$

Finalement, la tension musculaire $\tau$ est obtenue en superposant de façon asynchrone les contractions élémentaires (1) dues à $N_u$ unités motrices. Les positions $t_{j,k}$ des contractions élémentaires de l'unité motrice $k = 1, N_u$ sont les suivantes.

$$t_{j,k} = \mathcal{N}(j \times \mu, \nu \times \mu) + \mathcal{N}_k(0, \mu) + T_{l,k} \tag{3}$$

La distribution normale $\mathcal{N}(j \times \mu, \nu \times \mu)$ positionne les contractions musculaires dues à un neurone moteur et la distribution $\mathcal{N}_k(0, \mu)$ décale aléatoirement les positions des contractions qui sont dues à des unités motrices différentes. Dans le modèle de Titze, $\mathcal{N}$ et $\mathcal{N}_k$ sont les seuls sources de variabilité qui sont toujours présentes. Titze a exploré, en option, des variations inter-unités du taux d'émission $\lambda$ et de l'amplitude de la contraction élémentaire, sans découvrir des différences qualitatives avec le cas homogène.
La fréquence vocale non perturbée $\bar{F}_o$ est fixée par l'expérimentateur et la tension $\bar{\tau}$ est la moyenne des superpositions spatiales et temporelles des contractions musculaires (1). La fréquence instantanée $f_o$ et la racine carrée de la tension musculaire instantanée $\sqrt{\tau}$ sont supposées être proportionnelles (Titze, 1991). Par conséquent, la relation entre perturbations instantanées $f_o - \bar{F}_o$ et la tension instantanée $\tau$ est la suivante.

$$f_o - \bar{F}_o = \bar{F}_o \times (\sqrt{\tau/\bar{\tau}} - 1) \tag{4}$$

(Titze, 1991) utilisait le développement de Taylor de la relation (4). La présence du quotient $\tau/\bar{\tau}$ dans la formule des perturbations instantanées (4) rend superflue la modélisation des amplitudes absolues des contractions (1).
Le modèle "$\mathcal{N} - P$" suscite des questions théoriques et numériques qui sont discutées ici. Les réponses possibles sont à la base du modèle "$\mathcal{G} - I$" qui est expliqué ci-après.

$(i)$ Le modèle "$\mathcal{N}-P$" est encombrant d'un point de vue numérique. Il exige le placement, la copie et l'addition de milliers d'exemplaires de la contraction (1) pour une durée de simulation d'une seconde. Aussi, la simulation des potentiels d'action exige une modélisation numérique distincte de celle des secousses. Les temps $t_{j,k}-T_l$ dans (3) fixent alors les positions des impulsions unitaires qui simulent les potentiels d'action.

$(ii)$ La relation (2) montre que le modèle est non-physique en principe et on constate qu'il l'est en pratique dès que le coefficient de variation $\nu=\sigma/\mu$ des durées des intervalles $\Delta t_{iii}$ entre secousses est $>0.2$. A partir de cette valeur, on observe des durées négatives, c.-à-d. on constate que la $(n+1)_{ieme}$ secousse musculaire peut précèder la $n_{ieme}$. Ces erreurs ne peuvent pas être interceptées car la modélisation repose sur les positions absolues des secousses.

$(iii)$ L'utilisation obligée du coefficient de variation $\nu$ comme paramètre de contrôle implique que le taux d'émission $\lambda$ et la dispersion $\sigma$ des durées inter-impulsions $\Delta t_{iii}$ d'un même neurone ne sont pas indépendants. L'un diminue lorsque l'autre augmente.

$(iv)$ La distribution Gaussienne $\mathcal{N}$ dans (2) et (3) est problématique pour une autre raison. Elle tolère des intervalles inter-impulsions $\Delta t_{iii}$ plus courts que le temps de réfraction $T_{refr}$ qui devrait empêcher un neurone d'émettre deux impulsions successives arbitrairement proches.

$(v)$ (Titze, 1991) compare la contraction simulée (1) à des contractions observées et constate que la durée de la première dépasse de $30\%$ voire $50\%$ la durée des secondes. Cette observation soulève la question à savoir si la forme de la contraction influence les perturbations de $Fo$ observées ainsi que le problème de la prolongation non-nécessaire du temps de calcul suite à l'utilisation de contractions simulées extra-longues.

$(vi)$ L'obtention des valeurs des indices de perturbations de la fréquence vocale $F_o$ est obscure (Titze, 1991). Il semble qu'elle implique une moyenne des fréquences instantanées $f_o$ sur une durée fixe.

# 3 Le modèle "$\mathcal{G}-I$" des sources neurologiques des apériodicités vocales

Le modèle "$\mathcal{G}-I$" est inspiré du modèle "$\mathcal{N}-P$", mais évite les hypothèses *ad hoc* de celui-ci. Il est appelé "$\mathcal{G}-I$" ici parce qu'il repose sur le tirage au sort à l'aide d'une distribution $Gamma\ (\mathcal{G})$ des durées des intervalles $(I)$ entre potentiels d'actions simulés.

$(i)$ Modélisation de la contraction élémentaire musculaire.
Nous avons utilisé le modèle (1) que nous comparons à une contraction triangulaire dont le temps de descente du maximum à zéro est le double du temps de monté de zéro au maximum. A temps de montée égal, la durée $(T_{1\%})$ du modèle exponentiel (1) est typiquement $2.5\times$la durée du modèle triangulaire qui épouse mieux les formes observées et qui permet de réduire le temps de calcul.

$(ii)$ Superposition versus convolution.
Le placement à l'instant $t_i$ de la contraction revient à décaler $s(t)$ en la recopiant. Une alternative est la convolution de $s(t)$ avec une impulsion unitaire $\delta(t - t_i)$. La superposition des contractions élémentaires est numériquement équivalente à la simulation des potentiels d'action par une superposition d'impulsions unitaires suivie d'une convolution avec la contraction $s(t)$.

$(iii)$ Simulation des durées des intervalles inter-impulsions.
La simulation des potentiels d'action d'un neurone moteur repose sur le tirage au sort des durées des intervalles inter-impulsions $\Delta t_{iii}$. Deger (2012) montre que la distribution $\mathcal{G}(k, b)$ imite le mieux la distribution observée des $\Delta t_{iii}$ d'un neurone. Les deux paramètres $k$ et $b$ de la distribution sont liés à la moyenne $\mu = 1/\lambda$ et au coefficient de variation $\nu$ des durées $\Delta t_{iii}$ de la manière suivante.

$$k = 1/\nu^2, b = 1/\mu\nu^2 \tag{5}$$

La contrainte physique $\Delta t_{iii} \geq 0$ est satisfaite quelque soit la valeur des paramètres. La contrainte physiologique $\Delta t_{iii} > T_{refr}$ implique la condition $\nu < 1$ qui est nécessaire mais pas suffisante. C'est pourquoi, les durées $\Delta t_{iii}$ tirées au hasard et $< T_{refr}$ sont omises.

$(iv)$ Calcul des durées de cycles.
La fréquence instantanée $f_o$ est obtenue via la relation (4). Ensuite, les durées de cycle $T_o$ sont obtenues en cumulant le temps qui est nécessaire à la phase $\phi$ de croître de $2\pi$ sur base de la relation canonique $d\phi = 2\pi \times f_o \times dt$. Nous utilisons cette relation pour obtenir les durées de cycles simulées pour les modèles "$\mathcal{N} - P$" et "$\mathcal{G} - I$".

# 4 Méthodes

$(i)$ Nous avons calculé l'indice $J_{ppq5}$ des perturbations des durées de cycles afin de faciliter la comparaison avec des données publiées (Boersma & Weeninck, 2014). Nous utilisons une variante de $J_{ppq5}$ qui remplace la moyenne courante sur $5$ cycles par une moyenne courante sur un nombre de cycles qui est équivalent à une durée de $50ms$ (c.-à-d. $5$ cycles lorsque $Fo = 100Hz$). L'avantage en est que la durée moyenne des cycles n'est pas impliquée explicitement dans la définition du jitter vocal.

$(ii)$ La valeur de indice $J_{ppq5}$ ou d'autres qui lui sont similaires n'est pas suffisante afin de comparer les perturbations simulées et naturelles. En effet, ces indices rapportent des perturbations à court terme dont les valeurs peuvent êtres très semblables pour des scintillements lents et larges ou rapides et petits et il n'existe pas de données publiées qui permettraient de trancher.

$(iii)$ Nous avons estimé le scintillement vocal à l'aide de indice $S_{10}$ qui rapporte les perturbations par rapport une moyenne courante de $100ms$ des durées de cycles desquelles le $jitter$ a été soustrait. La moyenne courante sur $100ms$ est conforme à la définition du scintillement vocal (Titze, 1994). On observe que le quotient $J_{ppq5}/S_{10}$ découvre des différences entre options de modélisation qui échappent aux indices de $jitter$ conventionnels (cf. Tableau 1, colonne $9$ comparée aux colonnes $4$ et $7$).

$(iv)$ Afin d'offrir un point d'ancrage à la discussion des résultats de simulations, nous avons mesuré l'indice $J_{ppq5}$ du jitter et l'indice $S_{10}$ du scintillement des longueurs de cycles pour un corpus de

voyelles [a] soutenues par 35 et 59 locuteurs dont la voix a été libellée respectivement $G = 0$ ou $R = 0$ sur les échelles $GRBAS$ (U.MA., 2018). L'analyse porte sur un intervalle de $1sec$ placé $1sec$ après l'attaque de la voyelle échantillonnée à $44kHz$. Le Tableau 1 rapporte les quartiles des indices.

$(v)$ Les colonnes à l'extrême droite du tableau 1 rapportent les valeurs des mêmes indices pour une série de 1000 simulations des perturbations par du bruit blanc Gaussien des durées de cycles avec des valeurs aléatoires pour la fréquence vocale $\bar{F}o$ $(100\ Hz - 200\ Hz)$ et du coefficient de variation $\nu_{Fo}$ $(0.1\% - 0.5\%)$.
Un test non-paramétrique de Kolmogorov-Smirnov indique que les indices pour les corpus "$R = 0$" et "$G = 0$" ne diffèrent pas statistiquement. Par contre, les quotients diffèrent statistiquement significativement entre les colonnes 4 et 9 ainsi que 7 et 9 ($p < 10^{-3}$).

| | Corpus "G=0" | | | Corpus "R=0" | | | Bruit blanc | |
|---|---|---|---|---|---|---|---|---|
| Q | $S_{10}$ | $J_{ppq5}$ | $J_{ppq5}/S_{10}$ | $S_{10}$ | $J_{ppq5}$ | $J_{ppq5}/S_{10}$ | $J_{ppq5}$ | $J_{ppq5}/S_{10}$ |
| 05% | 0.08 | 0.12 | 0.76 | 0.08 | 0.12 | 0.81 | 0.09 | 2.58 |
| 25% | 0.12 | 0.16 | 1.06 | 0.11 | 0.19 | 1.23 | 0.15 | 3.30 |
| 50% | 0.15 | 0.20 | 1.26 | 0.15 | 0.24 | 1.53 | 0.22 | 3.92 |
| 75% | 0.19 | 0.26 | 1.72 | 0.20 | 0.29 | 2.25 | 0.29 | 4.42 |
| 95% | 0.30 | 0.39 | 2.47 | 0.34 | 0.48 | 3.25 | 0.36 | 5.17 |

Tableau 1 : Quantiles des indices du jitter ($J_{ppq5}$ en %) et scintillement ($S_{10}$ en %) vocal pour deux corpus de voyelles [a]. Colonnes à l'extrême droite : simulation des perturbations par du bruit blanc Gaussien des durées de cycles.

$(vi)$ Deux séries de 1000 simulations chacune ont été réalisées afin d'étudier les modèles "$\mathcal{G} - I$" et "$\mathcal{N} - P$" et de les comparer. Les modèles sont implémentés dans $Python$. La fréquence d'échantillonnage est à 200 $kHz$, la durée d'une simulation est une $sec$ et la fréquence vocale moyenne $\bar{F}_o$ est à 100 $Hz$ recto tono. Au début de chaque simulation, les valeurs des paramètres de contrôle sont choisies au hasard dans un intervalle dont les étendues relatives sont identiques afin de faciliter la comparaison entre paramètres (Tableau 2).

| | Paramètre | Valeur typique | Intervalle | Références |
|---|---|---|---|---|
| Temps mort | $T_{refr}$ | 2.5 ms | 1.25-3.75 ms | (Roark *et al.*, 2002) |
| Taux d'émission | $\lambda$ | 30 Hz | 15-45 Hz | (Roark *et al.*, 2002) |
| Ecart-type$\times\lambda$ | $\nu$ | 0.1 | 0.05-0.15 | (Titze, 1991) |
| Nombre U.M. | $N_u$ | 100 | 50-150 | (Titze, 1991) |
| Latence | $T_l$ | 15 ms | 7.5-22.5 ms | (Titze, 1991) |

Tableau 2 : Valeurs typiques et intervalles des paramètres de modélisation.

En outre, la contraction triangulaire ou la contraction exponentielle (1) a été choisie au hasard pour chaque simulation $(p = 1/2)$. Tous les paramètres sont identiques pour toutes les unités motrices qui sont co-actives lors d'une simulation. Le temps de montée $T_m$ est fixé à 0.02 $sec$ pour toutes les unités motrices et toutes les simulations. Aucun effort n'a été fait pour ajuster les modèles aux indices observés pour des locuteurs.

$(vii)$ La contribution relative de chaque paramètre est quantifiée à l'aide d'une analyse par régression linéaire multiple des variables z-normalisées. L'inverse $S_{10}^{-1}$ et $J_{ppq5}^{-1}$ des indices est utilisé comme

variable dépendante afin de mieux linéariser le lien entre variables. Tous les tests statistiques sont réalisés à l'aide du logiciel $R$.

# 5 Simulations et discussion

$(i)$ Une série préliminaire de 100 simulations avait comme objectif la comparaison de deux versions du modèle "$\mathcal{N}-P$". La version originale somme les contractions élémentaires directement et la version alternative convolue les potentiels d'action simulés avec la contraction $s(t)$. Les tensions musculaires calculées sont identiques pour les deux versions (corrélation $\equiv 1$), mais les temps de calcul diffèrent approximativement d'un ordre de grandeur. C'est pourquoi, toutes les simulations qui sont rapportées ci-après font appel à la convolution. Le gain en temps de calcul découle de l'implémention efficace de la convolution dans la bibliothèque $numpy$ et ne peut pas être généralisé à d'autres logiciels.

$(ii)$ Une série de 1000 simulations chacune a été réalisée avec les modèles "$\mathcal{G}-I$" et "$\mathcal{N}-P$". Le Tableau 3 montre les quartiles des indices. Un test de Kolmogorov-Smirnov indique que le quotient $J_{ppq5}/S_{10}$ diffère statistiquement significativement entre "$\mathcal{N}-P$" et "$\mathcal{G}-I$" ($p < 10^{-3}$).

| | "$\mathcal{G}-I$" | | | "$\mathcal{N}-P$" | | |
|---|---|---|---|---|---|---|
| Q | $S_{10}(\%)$ | $J_{ppq5}(\%)$ | $J_{ppq5}/S_{10}$ | $S_{10}(\%)$ | $J_{ppq5}(\%)$ | $J_{ppq5}/S_{10}$ |
| 05% | 0.13 | 0.16 | 0.74 | 0.10 | 0.17 | 1.07 |
| 25% | 0.21 | 0.29 | 1.23 | 0.18 | 0.31 | 1.38 |
| 50% | 0.29 | 0.52 | 1.79 | 0.28 | 0.53 | 1.75 |
| 75% | 0.40 | 0.84 | 2.44 | 0.45 | 0.87 | 2.36 |
| 95% | 0.75 | 1.75 | 3.76 | 0.89 | 1.67 | 3.71 |

Tableau 3 : Quantiles des indices obtenus pour deux modèles du jitter vocal.

$(iii)$ Le Tableau 4 rapporte les coefficients de la régression des indices $S_{10}^{-1}$, $J_{ppq5}^{-1}$ et $J_{ppq5}/S_{10}$ sur les paramètres des modèles. Le paramètre $\mathbb{C}$ réfère au type de la contraction (triangulaire $=$ "1", exponentielle $=$ "0"). La dernière ligne rapporte la qualité du modèle de régression. Les coefficients de régression dont seulement l'ordre de grandeur est donné sont statistiquement non-significatifs.

| | "$\mathcal{G}-I$" | | | "$\mathcal{N}-P$" | | |
|---|---|---|---|---|---|---|
| Paramètres | $S_{10}^{-1}$ | $J_{ppq5}^{-1}$ | $J_{ppq5}/S_{10}$ | $S_{10}^{-1}$ | $J_{ppq5}^{-1}$ | $J_{ppq5}/S_{10}$ |
| $T_l$ | $10^{-2}$ | $10^{-2}$ | $10^{-2}$ | $10^{-2}$ | $10^{-2}$ | $10^{-3}$ |
| $T_{refr}$ | $10^{-3}$ | $10^{-2}$ | $10^{-3}$ | $10^{-3}$ | $10^{-3}$ | $10^{-3}$ |
| $\lambda$ | $+0.52$ | $+0.81$ | $-0.55$ | $+0.67$ | $+0.70$ | $-0.15$ |
| $N_u$ | $+0.31$ | $+0.19$ | $10^{-3}$ | $+0.24$ | $+0.23$ | $-0.06$ |
| $\nu$ | $-0.57$ | $-0.64$ | $-0.50$ | $-0.40$ | $-0.15$ | $-0.51$ |
| $\mathbb{C}$ | $-0.28$ | $-0.27$ | $-0.13$ | $-0.35$ | $-0.33$ | $-0.06$ |
| $R^2$ ajusté | *0.77* | *0.75* | *0.56* | *0.82* | *0.69* | *0.29* |

Tableau 4 : Coefficients de régression linéaire multiples des paramètres z-normalisées.

$(iv)$ L'utilisation du coefficient de variation $\nu$ qui dépend du taux d'émission $\lambda$ via $\nu = \sigma \times \lambda$ facilite le contrôle de la causalité du modèle "$\mathcal{N}-P$" et a été retenue pour le modèle "$\mathcal{G}-I$" à des fins de

comparaison.
En inspectant les coefficients de régression (Tableau $4$), on constate que les indices de perturbations $S_{10}$ et $J_{ppq5}$ augmentent toujours avec la dispersion $\sigma$ des durées inter-impulsions et diminuent avec le taux d'émission $\lambda$ à condition que la dispersion est faible.
Les indices de perturbation augmentent également lorsque le nombre d'unités motrices diminue et lorsque la secousse triangulaire remplace la secousse exponentielle (1). Une explication possible est que la première est plus courte et moins lisse que la deuxième.

$(v)$ Le quotient $J_{ppq5}/S_{10}$ diminue lorsque la dispersion $\sigma$ ou le taux d'émission $\lambda$ augmentent. En d'autres mots, l'observation de valeurs $> 1$ du quotient implique que les rafales d'impulsions des neurones moteurs sont pseudo-périodiques. Pour rappel, la taille du jitter vocal est supérieure à la taille du scintillement vocal pour la parole naturelle (cf. Tableau 1). Titze (1991) ne discute pas le rapport entre jitter et scintillement vocal car l'hypothèse que les potentiels d'action sont pseudo-périodiques est implicite au modèle "$\mathcal{N} - P$". En effet, la condition $\nu$«$1$ est garante de la causalité du modèle. Le quotient $J_{ppq5}/S_{10}$ augmente aussi lorsque la contraction exponentielle remplace la contraction triangulaire, mais l'impact est faible.

$(vi)$ Le temps de latence de la contraction et le temps de réfraction du neurone moteur n'ont pas d'influence significative. Une explication possible est que le temps de latence est masqué par le terme aléatoire $\mathcal{N}_k$ dans la relation (3) et qu'il est peu probable qu'une durée inter-impulsions $\Delta t_{iii}$ aussi petite que le temps de réfraction soit observée lorsque le coefficient de variation $\nu$ des durées $\Delta t_{iii}$ est $< 0.15$.

$(vii)$ Les valeurs des indices $S_{10}$ et $J_{ppq5}$ ont les mêmes ordres de grandeur pour les simulations et les observations sur des locuteurs (tableaux 1 et 3), mais leurs étendues diffèrent. En effet, rien n'a été fait pour ajuster les paramètres de manière à ce que les modèles reproduisent les valeurs observées. Aussi, (Titze, 1991) ne discute pas la possibilité que les perturbations des fréquences instantanées $f_o$ soient modulées à l'intérieur des plis et que les perturbations qui sont simulées par "$N - P$" ou "$G - I$" ne soient pas observables directement. En effet, des conditions laryngées (bénignes) existent qui sont connues influencer le jitter vocal mesuré sans pour cela influencer l'activité des unités motrices. Il se peut, qu'une explication possible est inhérente au modèle $muscle - couverture$ des plis vocaux. Un modèle simple qui n'est pas reproduit ici faute de place, mais qui tient compte des amplitudes de vibration du muscle $A_m$ et de la couverture $A_c$ des plis, suggère que les perturbations instantanées $f_o - \bar{F}_o$ sont multipliées par un terme $A_m/(A_m + A_c)$. Ce terme est égal à l'unité lorsque $A_c = 0$ et égal à zéro lorsque $A_m = 0$. Il suggère donc que les perturbations instantanées sont modulées par la mobilité relative du muscle et de la couverture.

# Références


ALZAMENDI G. H. & SCHLOTTHAUER G. (2017). *Describing Voice Period Variability by means of Time Series Structural Analysis*. Proceedings 10th International Workshop : Models and Analysis of Vocal Emissions for Biomedical Applications, Firence, Italy, pages 11-14.

BOERSMA P. & WEENINCK D. (2014). *Praat : doing phonetics by computer [Computer program]*. [Version 5.4.04, retrieved 2014 from http ://www.praat.org].

P. R. COOK, Ed. (1999). *Music, Cognition and Computerized Sound :An Introduction to Psychoacoustics*. Cambridge, Massachusetts : The MIT Press. page 199.

DEGER M., HELIAS M., BOUCSEIN C. & ROTTER S. (2012). Statistical properties of superimposed stationary spike trains. *J. Comput. Neurosci.*, **32**, 443–463.

HERMAN I. P. (2007). *Physics of the Human Body*. Berlin, Heidelberg : Springer. page 281.

KREIMAN J. & SIDTIS D. (2011). *Foundations of Voice Studies : An Interdisciplinary Approach to Voice Production and Perception*. Wiley-Blackwell. page 55.

ROARK R., C.L., LI J., SCHAEFER S., ADAM A. & LUCA C. D. (2002). Multiple motor unit recordings of laryngeal muscles : The technique of vector laryngeal electromyography. *The Laryngoscope*, **112**, 2196–2202.

TITZE I. R. (1991). A model for neurologic sources of aperiodicity in vocal fold vibration. *J. Speech, Hearing Res.*, **34**, 460–472.

TITZE I. R. (1994). *Principles of Voice Production*. Englewood Cliffs, NJ : Prentice Hall. page 332.

U.MA. (2018). *http ://www.atic.uma.es/index_atic.html*. [Online ; accessed 23-January-2018].

# Un algorithme de segmentation en phrasé

Philippe Martin[1]

(1) LLF, UFRL, ODG, Place Paul Ricoeur, 75013 Paris, France
philippe.martin@linguist.univ-paris-diderot.fr

## RESUME

En lisant à voix haute ou silencieusement, nous segmentons la parole en groupes accentuels, ne contenant qu'une seule syllabe accentuée (hors accent d'insistance). Dans les langues à accent lexical comme l'anglais, les groupes accentuels contiennent un mot lexical (nom, adverbe, verbe ou adjectif) dans lequel la position de l'accent est définie dans le lexique. En français, langue sans accent lexical, les groupes accentuels sont définis non par la catégorie de mots qu'ils contiennent mais par le temps qu'il faut pour les lire ou les prononcer. Le phrasé, c'est-à-dire la segmentation en groupes accentuels, dépend donc du débit de parole. Avec un débit de parole lent, toutes les syllabes finales de mots lexicaux sont accentuées, alors qu'un débit de parole rapide peut rassembler plusieurs mots lexicaux dans un seul groupe accentuel. Un algorithme de phrasé basé sur cette propriété est présenté et illustré sur deux exemples de parole lue et spontanée.

## ABSTRACT

**Automatic phrasing in French.**
Whether we read aloud or silently, we segment speech in accent phrases, containing only one stressed syllable (excluding emphatic stress). In lexically stressed languages such as English, the location of stress in a noun, an adverb, a verb or an adjective (content words) is defined in the lexicon, and accent phrases include one single content word. In French, a language without lexical stress, accent phrases are defined by the time it takes to read or pronounce them. Therefore, actual phrasing, i.e. the segmentation into accent phrases, depends on the speech rate. With a slow speech rate, all content words final syllables are stressed, whereas a fast speech rate could merge more than one content word in a single accent phrase. Based on this observation, a computer algorithm for automatic phrasing operating in a top-down fashion is presented and applied to two examples of read and spontaneous speech.

MOTS-CLES : Groupe accentuel, français, accent lexical, phrasé, syllabe accentuée
KEYWORDS: Accent phrase, French, lexical stress, phrasing, syllabic stress

# 1 Introduction

Lorsque nous lisons un texte en anglais, à voix haute ou silencieusement, nous pouvons procéder mot à mot ou même syllabe par syllabe, mais si nous maîtrisons la langue et identifions tous les mots, nous procédons généralement par groupe de mots. Il est facile d'observer dans une


10.21437/JEP.2018-32

transcription orthographique où tous les mots seraient terminés par un point final que nous ne lisons pas mot à mot, comme ce serait le cas dans l'exemple en anglais : *In. The. Orthographic. Representation. Of. Speech. Of. Most. Written. Languages. Segmentation. Is. Defined. By. Spaces. Between. Words.* En réalité, si nous sommes suffisamment familiers avec la langue en question, nous lisons normalement en regroupant les mots en unités contenant un nom, un adverbe, un verbe ou un adjectif (i.e. un mot lexical, de classe ouverte), accompagnés chacun par des mots grammaticaux (pronoms, conjonctions, prépositions, déterminants..., mots de classe fermée) qui leur sont associés : [*in the ortho**gra**phic*] [*represen**ta**tion*] [*of **speech***] [*of **most***] [***wri**tten*] [***lan**guages*] [*segmen**ta**tion*] [*is de**fi**ned*] [*by **spa**ces*] [*bet**ween***] [***words***]. En phonologie, de tels groupes de mots sont appelés groupes accentuels, et définissent les unités prosodiques minimales, qui organisés en une hiérarchie, constituent la structure prosodique de la phrase (Martin, 1975, Selkirk, 1978).

Pour tous les locuteurs de l'anglais, la position des syllabes accentuées dans les mots lexicaux est prévisible et résulte de l'acquisition du lexique de la langue. D'autres syllabes accentuées peuvent également apparaître, mais contrairement à l'accent lexical, elles ne sont pas prévisibles car elles résultent d'un choix particulier du locuteur pour indiquer une emphase. Ce type d'accentuation emphatique peut se produire sur une syllabe différente de l'accent lexical ou sur la même syllabe. Dans ce dernier cas, le locuteur utilisera une réalisation acoustique différente car l'accent emphatique doit être perçu par les auditeurs comme différent et imprévisible par rapport à l'accent lexical prévisible.

La prévisibilité de l'accent lexical suggère que la perception des syllabes accentuées ne dérive pas directement du traitement des caractéristiques acoustiques spécifiques du discours, telles que la durée de la voyelle, le changement de fréquence fondamentale ou la modulation d'intensité, les paramètres prosodiques classiques souvent mentionnés dans la littérature comme paramètres de l'accent. En fait, la perception des syllabes accentuées peut être considérée comme le résultat d'un mécanisme d'identification comparant les caractéristiques acoustiques réelles des syllabes avec une position prévisible dérivée de la connaissance de la langue.

On peut citer à ce sujet l'expérience sur la perception des syllabes accentuées du berbère et de l'hébreu par des sujets qui n'ont aucune notion de ces langues (Mettouchi et al., 2007). Les caractéristiques acoustiques de l'accent syllabique sont présentes dans le signal vocal, mais dans cette expérience, les auditeurs n'ont identifié que très peu de positions correctes de l'accent lexical, puisqu'ils ne disposaient d'aucun lexique approprié permettant de prépositionner les syllabes accentuées lors de l'audition des mots correspondants, contrairement aux locuteurs du berbère ou de l'hébreux.

# 2 Les groupes accentuels en français

Le français est une langue où la position de l'accent lexical a évolué progressivement vers la dernière syllabe des mots lexicaux (et même sur la dernière syllabe de tous les mots prononcés isolément) en perdant progressivement toutes les syllabes placées à l'origine après l'accent (Väänänen, 1995). La fonction de l'accent lexical comme marqueur de frontière morphologique existant dans les autres langues romanes a été progressivement perdue puisque devenue redondante. Il est alors devenu possible pour les locuteurs d'ignorer certains accents syllabiques portés par les mots lexicaux, comme dans *la petite armoire violette*, qui peut recevoir une, deux ou même trois syllabes accentuées : *la petite armoire vio**lett**e, la petite ar**moir**e vio**lett**e , la pe**tit**e armoire vio**lett**e* ou *la pe**tit**e ar**moir**e vio**lett**e*. Pour un locuteur francophone, il est facile de se rendre compte que la différence d'accentuation de ces exemples est liée au débit de parole. Pour

prononcer (ou même lire silencieusement) *la petite armoire violette* avec une seule syllabe finale accentuée sur *vio**lette***, il faut utiliser un débit de parole (très) rapide, alors qu'un débit plus lent conduit à la prononciation de trois syllabes accentuées dans le même exemple.

On pourrait peut-être conclure qu'il n'y a pas de limite au nombre de syllabes et donc de mots qui peuvent être prononcés en français avec une seule syllabe accentuée finale, et qui peut être contenue dans un seul groupe accentuel. La prononciation des mots longs permet toutefois de déterminer une limite. Des mots tels que *l'anticonstitutionnalité* ("contre la constitution"), (8 syllabes) ou *intergouvernementalisation* ("inter-gouvernemental") (10 syllabes) semblent difficiles voire impossibles à prononcer ou même à lire silencieusement avec une seule syllabe accentuée finale. Déjà au XVème siècle, le grammairien Louis Meigret (1550) concluait que le mot le plus long qui pouvait être prononcé avec un seul accent final est formé d'un maximum de 7 syllabes. Bien plus tard, Martin (2014) a montré que ce n'est pas le nombre de syllabes qui importe, mais le temps qu'il faut pour les prononcer. Les données expérimentales montrent en effet que l'intervalle maximum entre deux syllabes accentuées consécutives (dans la parole continue) ne peut pas dépasser 1250 ms environ. Dans le style « parole de jeunes », on trouve des séquences de 10 ou 11 syllabes avec seulement une seule syllabe accentuée finale. Cette valeur est proche de la limite théorique, dérivée de la durée moyenne minimale des syllabes qui pourraient être perçues dans une séquence, soit 100 ms (Ghitza & Greenberg, 2009). Ces observations situent la durée maximale des phrases accentuées en français à environ 1250 ms à 1400 ms, le débit le plus rapide atteignant 8 à 9 syllabes par seconde (Lekha et Le Gac, 2004).

On peut aussi établir une durée minimale séparant deux syllabes accentuées successives, dans une configuration dite de « collision accentuelle ». Sa valeur s'évalue en sélectionnant des occurrences telles que *par le **fait** **que*** ou *le travail de **nuit** **nuit***, sans déplacement ou suppression possible du premier accent. Il est souvent mentionné dans la littérature que ces cas nécessitent un écart acoustique entre syllabes accentuées consécutives (par exemple Di Cristo, 2016), habituellement mis en œuvre par la présence de consonnes après la première syllabe accentuée. En réduisant progressivement avec un éditeur de signal l'écart jusqu'à ce que la première syllabe cesse d'être perçue comme accentuée (sans modifier leur structure acoustique, i.e. en supprimant seulement la partie silencieuse), on obtient une limite d'environ 250 ms, ce qui détermine la durée minimale d'un groupe accentuel constitué de la deuxième syllabe de la séquence et précédé d'un silence suffisant. La désaccentuation perçue de la première syllabe [*par le **fait***] va alors restructurer le groupe accentuel en [*par le fait **que***].

# 3 Syllabes finales

Il est également facile de démontrer expérimentalement que toute syllabe suivie d'au moins 250 ms de silence est perçue comme accentuée en français. En insérant un silence de 250 ms, les syllabes finales de n'importe quelle catégorie de mots sont perçues comme accentuées, quelle que soit leur durée réelle ou leur mouvement de hauteur mélodique. Dans les langues à accent lexical, la perception d'une syllabe finale d'accent comme accentuée est préemptée par la position de l'accent lexical (s'il n'est pas en position finale). En italien par exemple, l'accent lexical de l'avant-dernière syllabe de *Marco* dans *la So**re**lla di **Mar**co à par**ti**ta* ("la sœur de Marco est partie") empêche un auditeur connaissant la langue de percevoir la dernière syllabe de *Marco* comme accentuée, même si elle est réalisée avec une forte montée mélodique, alors que pour un locuteur de français ne connaissant pas l'italien, cette syllabe sera probablement perçue comme accentuée, puisque c'est sa position attendue.

# 4 Glissando

Selon le modèle de Martin (1975, 2018), la structure prosodique résulte d'une organisation hiérarchique des groupes accentuels. En référence à un contour terminal attendu, perçu comme un marqueur de non-continuation de la phrase, deux autres contours mélodiques, les uns montants, les autres descendants, indiquent respectivement une continuité majeure et une continuité mineure (pour reprendre la terminologie déjà proposée par Delattre en 1966, mais avec une autre définition puisque portant ici sur les seules voyelles accentuées).

Les contours de continuité indiquent une relation de dépendance, de la continuation mineure vers la continuation majeure et de la continuation majeure vers le contour terminal, par un contraste de pente mélodique, où un contour descendant indique une dépendance en vers un contour montant situé plus loin dans le temps. Ce modèle implique que les mouvements mélodiques descendants et montants soient effectivement perçus comme tels, c'est-à-dire que la vitesse du changement mélodique dans le temps soit supérieure au seuil de glissando. Ce seuil est évalué à partir de la différence entre le début et la fin de la variation mélodique évaluée en demi-tons par rapport à la durée du contour (en supposant une variation linéaire, voir Rossi, 1971). Toute syllabe de mot portant une variation mélodique supérieure au seuil de glissando est donc accentuée selon le modèle de la structure prosodique retenu.

# 5 Pronoms toniques et démonstratifs

Les pronoms toniques en français (*moi, toi, lui, elle, nous, vous, eux, elles)* n'appartiennent pas à la catégorie des mots lexicaux, mais partagent leurs caractéristiques en termes d'accentuation, en particulier en position postverbale. En particulier, ils seront accentués s'ils sont suivis d'au moins 250 ms de silence ou si leur variation mélodique est supérieure au seuil de glissando. Ainsi dans l'exemple *moi ma mère le salon c'est de la moquette*, le pronom tonique *moi* est accentué s'il est suivi de 250 ms de silence, ***moi*** *# ma mère le sa****lon*** *c'est de la mo****quett****e*, mais ne l'est pas s'il n'y a pas de silence suffisant après *moi* : *moi ma* ***mèr****e* .... Il en va de même pour les pronoms démonstratifs (*celui, ceux, celles, celui-ci, etc.*) accentuables même non suivis d'un silence de de 250 ms.

# 6 Eurythmie

Selon Wioland (1985), l'eurythmie de la parole spontanée procède en ajustant la durée moyenne des syllabes accentuées pour atteindre une durée comparable des groupes accentuels successifs. En lecture, les locuteurs utilisent le plus souvent une stratégie visant à équilibrer le nombre de syllabes des groupes accentuels successifs, au détriment éventuel de la congruence avec la structure syntaxique. Un exemple classique est donné par la phrase *Marie adore les chocolats*, où le locuteur a tendance en parole spontanée à réaliser un phrasé congruent avec la syntaxe [*Ma****rie***] [*adore les choco****lat****s*] et éventuellement à atteindre l'eurythmie en ralentissant le débit syllabique de [*Ma****rie***] et en accélérant sur [*adore les choco****lat****s*]. Au contraire, les lecteurs de cette même phrase montrent une tendance à regrouper les mots pour équilibrer le nombre de syllabes en phrases d'accent consécutives, au détriment de la congruence avec la syntaxe [*Marie adore*] [*les chocolats*].

# 7 Syllabes accentuables et syllabes accentuées

Le phrasé détermine une étape essentielle dans la compréhension de la parole. La segmentation en groupes accentuels constitue la première phase de reconstruction de la structure prosodique voulue par le locuteur, indispensable et incontournable pour accéder à la structure syntaxique dans une étape ultérieure. La structure prosodique résultante ne correspond pas nécessairement à la structure prosodique voulue par l'auteur du texte lu, car le phrasé dépend de la vitesse de lecture choisie, à haute voix ou silencieusement.

Le simple fait que nous connaissons la position possible des syllabes accentuées lorsque nous lisons à haute voix ou silencieusement suggère que nous n'avons pas vraiment besoin de données acoustiques pour percevoir des syllabes accentuées (non-emphatiques). Non seulement la lecture à voix haute ou silencieuse du même texte peut conduire à des segmentations différentes, mais à l'écoute, on ne peut empêcher d'avoir des attentes vis-à-vis d'une localisation des syllabes accentuées différente de celle effectivement réalisée par le locuteur. En d'autres termes, nous pouvons "entendre" des syllabes accentuées qui peuvent ne pas être présentes acoustiquement. Cette illusion apparente se retrouve dans de nombreux processus impliqués dans la perception de la parole (Arnal & Giraud, 2017), et suggère non pas un traitement direct d'une entrée physique, mais la validation d'une entrée attendue par comparaison entre ce qui est attendu et ce qui est physiquement réalisé.

Puisque la réalisation effective de l'accentuation dépend de la vitesse de prononciation, la seule manière d'éviter la perception d'un accent syllabique virtuel qui ne serait pas réalisé effectivement serait d'adapter constamment le débit à celui utilisé par le locuteur. Cette adaptation n'est pas toujours facile ni même possible. Des exemples avec un débit de parole très rapide dépassant 7 ou 8 syllabes par seconde sont difficiles à suivre pour la plupart des auditeurs, au point que certains auront du mal à comprendre les énoncés, et avoir tendance à entendre des syllabes accentuées là où elles n'existent pas acoustiquement.

L'écart possible entre les syllabes accentuées perçues et celles effectivement réalisées en français conduit à différencier les syllabes accentuables des syllabes effectivement accentuées, selon la terminologie de Paul Garde (1968, 2013).

# 8 Annotation des syllabes accentuées : mission impossible ?

Le problème pour un annotateur de syllabes accentuées en français est donc de s'adapter au débit de parole de l'enregistrement. La perception sera influencée par le processus de prédiction de l'annotateur, tendant à détecter les syllabes accentuées aux endroits où il les aurait placées en lisant ou en parlant non pas avec le débit du locuteur mais avec le sien. On peut aussi imaginer pour la même raison qu'un annotateur ne perçoive pas une syllabe effectivement accentuée.

Le plus souvent, la détection automatique des syllabes accentuées en français opère de bas en haut (bottom-up) à partir de l'enregistrement, recherchant des variations acoustiques significatives entre syllabes consécutives en durée, fréquence fondamentale et intensité (pour les exemples récents, voir Goldman et al., 2013 ; Mertens & Simon, 2013). Dans ces processus, la qualité des voyelles n'apparaît pas comme un paramètre significatif en français.

Dans un article publié en 2013, M. Avanzi, confronté à l'incertitude dans l'annotation des syllabes accentuées, décrit en détail une procédure complexe impliquant deux experts, éventuellement aidés

d'un troisième en cas de désaccord entre les deux premiers. Même avec ce protocole, l'accord entre les annotateurs varie entre 60% et 80%.

Dans un autre article sur le même sujet, Christodoulides et Avanzi (2014) ont mis en œuvre un détecteur automatique de proéminence (c'est-à-dire également de l'accent d'emphase) par des méthodes d'apprentissage automatique appliquées à un grand corpus d'une durée de 11 heures. Les auteurs utilisent un ensemble complet de paramètres acoustiques censés être appropriés pour différencier les syllabes proéminentes des autres syllabes (durée syllabique minimale et maximale, fréquence fondamentale moyenne, minimale et maximale, intensité maximale, équilibre spectral, partie de l'étiquette vocale, présence et durée des pause, structure syllabique, position de la syllabe dans le mot, etc.). Leurs meilleurs résultats, évalués par rapport à la référence déterminée par des experts, atteignent un niveau d'identification correct de 90%.

Compte tenu de ces difficultés, il semble que la détection des syllabes accentuées ne devrait pas procéder directement de l'analyse du signal de parole, mais plutôt indirectement par validation des hypothèses relatives aux positions potentielles basées sur les observations faites plus haut.

# 9 Un algorithme de segmentation descendant (top-down)

Pour exploiter les données acoustiques, et innover par rapport aux approches *bottom-up* opérant à partir des données acoustiques, on propose ici un algorithme *top-bottom* basé sur les mécanismes cités plus haut, en retenant les règles suivantes :

1. Toute syllabe suivie d'un silence de plus de 250 ms est accentuée

2. Toute syllabe finale d'un nom, adjectif, verbe, adverbe ou pronom (tonique ou démonstratif) est accentuable (définition classique du groupe accentuel contenant un seul mot lexical)

3. Si deux syllabes accentuables ou accentuées successives sont séparées par moins de 250 ms, la première n'est pas accentuée (durée minimale du groupe accentuel)

4. Toute syllabe accentuable avec changement de F0 au-dessus du seuil de glissando est accentuée (définition de la structure prosodique)

5. Si 2 syllabes accentuées consécutives sont séparées de plus de 1250 ms en parole continue, au moins une syllabe accentuable dans cet intervalle est accentuée (durée maximale du groupe accentuel). Celle ayant avec la plus haute valeur de glissando est accentuée (caractère approximatif du calcul du glissando)

6. Une syllabe accentuable doit exister dans n'importe quelle durée de fenêtre égale à la durée moyenne de l'accentuation (eurythmie). L'eurythmie est mise en œuvre par le calcul incrémental de la moyenne des durées des groupes accentuels successifs à partir du début de l'enregistrement. Un test de cohérence eurythmique est ensuite appliqué par une fenêtre temporelle glissante, censée correspondre à un débit de parole supposé constant dans tout l'enregistrement. La durée de cette fenêtre résulte de la moyenne cumulative des groupes accentuels considérés successivement. Une syllabe accentuée supplémentaire sélectionnée selon sa valeur de glissando est ajoutée en cas d'absence d'accent dans une fenêtre donnée.

# 10 Un exemple de parole lue

Premier exemple de parole lue : *il était une fois un pauvre escargot qui souffrait beaucoup à chaque fois qu'il partait en randonnée car il avait du mal à suivre le rythme de ses compagnons*. Dans les étapes détaillées ci-dessous, les syllabes accentuables sont soulignées et les syllabes retenues comme effectivement accentuées sont soulignées et en gras.

1 : Toute syllabe finale de mot suivie d'un silence de plus de 250 ms est accentuée, cas du mot final *rando**nnée*** :
*Il était une fois un pauvre escargot qui souffrait beaucoup à chaque fois qu'il partait en rando**nnée***

2 : Toute syllabe finale d'un nom, adjectif, verbe, adverbe ou pronom est accentuable :
*Il était une fois un pauvre escargot qui souffrait beaucoup à chaque fois qu'il partait en rando**nnée***

3 : Si 2 syllabes accentuées successives sont séparées par moins de 250 ms, la première n'est pas accentuée : l'écart entre *chaque* et *fois* est de 180 ms, sous la limite de 250 ms, *chaque* ne peut être accentué :
*Il était une fois un pauvre escargot qui souffrait beaucoup à chaque* |**180 ms**| *fois qu'il partait en rando**nnée***

4 : Toute syllabe accentuable avec changement de F0 au-dessus du seuil de glissando est accentuée (notation {valeur de glissando / seuil de glissando avec coefficient 0,16} ). Les syllabes accentuables en dessous du seuil ne sont pas accentuées :
*Il était* {35/76} *une **fois*** {36/17} *un pauvre* {44/66} *escar**got*** {32/12} *qui souffrait* {54/144} *beau**coup*** {79/66} *à chaque fois* {46/106} *qu'il partait* {32/51} *en rando**nnée***

5 : Deux syllabes accentuées successives séparées de plus de 1250 ms, comme dans le cas de [*à chaque fois qu'il partait en randonnée*] de durée 1367 ms. En sélectionnant la valeur glissando la plus haute, sur *fois* : [*à chaque **fois** qu'il partait en rando**nnée***].

6 : Cohérence eurythmique. Les durées moyennes cumulées pour chaque groupe accentuel déterminé au stade précédent sont successivement : *il était une **fois*** 726 ms, *Il était une **fois** un pauvre escar**got*** 706 ms, *Il était une **fois** un pauvre escar**got** qui souffrait beau**coup*** 606 ms, le dernier groupe *qu'il partait en rando**nnée*** a une durée de 1033 ms, supérieure à la moyenne cumulée de 606 ms. La syllabe finale de *par**tait*** ayant la plus grande valeur de glissando est ajoutée à la liste des syllabes accentuées. Le résultat final est alors : *Il était une **fois** un pauvre escar**got** qui souffrait beau**coup** à chaque **fois** qu'il par**tait** en rando**nnée***.

# 11 Un exemple de parole spontanée

Le deuxième exemple présente un débit rapide caractéristique du « parler jeune » » : *Juste pour une carte d'identité tu n'as pas ta carte tu fais tes vingt-quatre heures tu en ressors t'as la haine encore plus ça augmente*.

1 : La dernière syllabe de l'énoncé est suivie de plus de 250 ms de silence :

*Juste pour une carte d'identité t'as pas ta carte tu fais tes vingt-quatre heures tu ressors t'as la haine encore **plus***

2 : Toute syllabe finale d'un nom, d'un adjectif, d'un verbe ou d'un adverbe est accentuable :
*Juste pour une carte d'identité t'as pas ta carte tu fais tes vingt-quatre heures tu ressors t'as la haine encore **plus***

3 : Si deux syllabes accentuables ou accentuées successives sont séparées par moins de 250 ms, la première n'est pas accentuée : les écarts entre *vingt-quatre* et *heures* (230 ms) et entre *encore* et ***plus*** (240 ms) sont en dessous de la limite de 250 ms :
*Juste pour une carte d'identité t'as pas ta carte tu fais tes vingt-quatre* | ***230 ms*** | *heures tu ressors t'as la haine encore* | ***240 ms*** | ***plus***

4: Toute syllabe accentuable avec changement de F0 au-dessus du seuil de glissando est accentuée. Les syllabes accentuables sous le seuil ne sont pas accentuées : **Juste** {64/36} pour une **carte** {44/38} d'identi**té** {54/45} t'as pas {18/35} ta **carte** {44/38} tu fais {54/142} tes vingt-quatre | 230 ms | **heur**es {49/37} tu re**ssor**s {38/32} t'as la **hain**e {25/22} encore | 240 ms | **plus** {38/23}
L'étape 5 ne s'applique pas, et le test d'eurythmie d'ajoute pas de syllabes accentuées supplémentaires.

# 12 Conclusion

Basé sur le fait que la perception des syllabes accentuées résulte d'un processus de validation comparant la position prédite avec des paramètres acoustiques effectivement mesurés, une segmentation automatique descendante du phrasé en français est décrite brièvement. L'algorithme incorpore les observations suivantes : 1) Durée minimale des groupes accentuels de 250 ms et maximale de 1250 ms ; 2) Toute syllabe finale de mot suivie d'au moins 250 ms de silence est perçue comme accentuée 3) la durée des groupes accentuels dépend du débit de parole choisi par le locuteur ou le lecteur ; 4) L'accent syllabique définissant le phrasé comporte un mouvement mélodique supérieur au seuil de glissando. Les données acoustiques n'interviennent que par les durées entre syllabes accentuables successives et la valeur de glissando de leur contour mélodique.

# Remerciements

À Damien Lolive, lecteur convaincant du conte « Le petit escargot »
(https://www.iletaitunehistoire.com/genres/albums-histoires/lire/la-maison-de-l-escargot-biblidhis_027).

# Références

ARNAL, L. et GIRAUD, A-L. (2017). Neurophysiologie de la perception de la parole et multisensorialité. *Traité de neurolinguistique*, Serge Pinto et Marc Sato éd., Louvain-la-Neuve : De Boeck, 97-108.

AVANZI, M., LACHERET-DUJOUR, A., and VICTORRI, B. (2010). A Corpus based Learning Method for Prominence Detection in Spontaneous Speech. *Proc. of Prosodic Prominence*, Speech Prosody Workshop.

AVANZI, M. (2013). Note de recherche sur l'accentuation et le phrasé à la lumière des corpus du français. *Tranel*, vol. 58, 5-24.

CHRISTODOULIDES, G. & AVANZI, M. (2014). An Evaluation of Machine Learning Methods for Prominence Detection in French. *Proc. Interspeech 2014*, 116-119.

DI CRISTO, A. (2016). *Les musiques du français parlé*, Berlin : De Gruyter Mouton. 513 p.

GARDE, P. (1968). *L'accent*. Paris : Presses universitaires de France, collection SUP « Le linguiste », n° 5. 172 p.

GARDE P. (2013). *L'accent*. Paris : Lambert-Lucas.

GOLDMAN, J-P., AUCHLIN, A., ROEKHAUT, S., SIMON, A-C., AVANZI, M. (2013). Prominence perception and accent detection in French. A corpus-based account, *Language Science* (39), 95-106.

GHITZA, O. & GREENBERG, S. (2009). On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, vol. 66, n$^{os}$ 1-2, 113-126.

LEHKA, I. & LE GAC, D. (2004). Etude d'un marqueur prosodique de l'accent de banlieue, *Actes des XXIIIème Journées d'Etudes sur la Parole*, avril 2004, Fès, Maroc.

MARTIN, P. (1975). Analyse phonologique de la phrase française. *Linguistics*, vol. 146, 35-68.

MARTIN, P. (2014). Spontaneous speech corpus data validates prosodic constraints. N. CAMPBELL, D. GIBBON, D. HIRST (éd.), *Proceedings of the 6$^{th}$ conference on speech prosody*, 525-529.

MARTIN, P. (2018). *Intonation, structure prosodique et ondes cérébrales*. London : ISTE.

MEIGRET, L. (1550). *Le tretté de la grammére françoéze*. Paris, C. Wechel.

MERTENS, P. & SIMON, A-C. (2013). Towards automatic detection of prosodic boundaries in spoken French, *Proceedings of the Prosody-Discourse Interface Conference 2013* (IDP-2013), 81-87.

METTOUCHI, A., LACHERET-DUJOUR, A., SILBER-VAROD, V. *et al.* (2007). Only Prosody ? Perception of speech segmentation in Kabyle and Hebrew. *Cahiers de linguistique française*, vol. 28, A. AUCHLIN (éd.), *Actes du 2$^{e}$ Symposium international IDP07 (Interfaces Discours Prosodie)*. 207-218.

ROSSI, M. (1971). Le seuil de glissando ou seuil de perception des variations tonales pour la parole. *Phonetica*. n° 23, 1-33.

SELKIRK, E. O. (1978). On prosodic structure and its relation to syntactic structure. In T. Fretheim, ed., *Nordic Prosody II*. Trondheim: TAPIR, 111-140.

VÄÄNÄNEN, V. (1995). *Introducción al latin vulgar*, Madrid: Editorial Gredos.

WIOLAND, F. (1985). *Les structures rythmiques du français*, Paris : Slatkine-Champion.

# Une mesure d'intelligibilité par décodage acoustico-phonétique de pseudo-mots dans le cas de parole atypique

Alain Ghio[1], Muriel Lalain[1], Laurence Giusti[1], Gilles Pouchoulin[1], Danièle Robert[1,2], Marie Rebourg[1], Corinne Fredouille[3], Imed Laaridh[3], Virginie Woisard[4]
(1) Aix-Marseille Univ, CNRS, LPL, UMR 7309, Aix-en-Provence, France
(2) Service ORL, APHM, Marseille, France
(3) Laboratoire d'Informatique d'Avignon, Avignon, France
(4) Service ORL, CHU Larrey, URI Octogone-Lordat, Toulouse, France

alain.ghio@lpl-aix.fr

## RESUME

Les limitations actuelles des tests d'intelligibilité effectués sur des locuteurs ayant une production atypique de la parole résident dans la capacité des auditeurs à restaurer les séquences distordues. Le résultat est une mesure surévaluée par rapport à la performance articulatoire réelle. Nous présentons un test d'intelligibilité fondé sur la prononciation de pseudo-mots de façon à complètement neutraliser les effets de lexicalité ou d'apprentissage des items par les auditeurs.
126 locuteurs (41 sujets sains et 85 patients atteints de troubles de la parole) ont produit chacun 52 pseudo-mots tirés aléatoirement d'une liste de 89346 formes possibles. 40 auditeurs ont retranscrit ces productions. Les transcriptions orthographiques ont été phonétisées puis comparées aux formes phonétiques attendues par un algorithme de Wagner-Fischer qui intègre les phénomènes d'insertion, élision et substitution de phonèmes. Les résultats montrent que les formes perçues chez les patients sont en moyenne à une distance bien plus élevée que chez les sujets contrôles.

MOTS-CLES : Intelligibilité; parole atypique; traits phonétiques ; décodage acoustico-phonétique

## ABSTRACT

**A measure of intelligibility by acoustic-phonetic decoding of pseudo-words in the case of atypical speech**

The current intelligibility tests performed on speakers with atypical speech production are limited by the ability of listeners to restore distorted sequences. The result is an overvalued measure compared to the actual articulatory performance. We present an intelligibility test based on the pronunciation of pseudo-words in order to neutralize unwanted lexical and learning effects of items by the listeners. 126 speakers (41 healthy subjects and 85 patients) each produced 52 pseudo-words randomly drawn from a list of 89346 possible forms. 40 listeners have transcribed these productions. Orthographic transcriptions were phonetized and compared to the phonetic forms expected by a Wagner-Fischer algorithm that integrates the phenomena of insertion, elision and phoneme substitution. The results show that the forms perceived with patients are on average at a greater distance than with healthy subjects.

KEYWORDS: Intelligibility; atypical speech; phonetic features ; acoustic-phonetic decoding


10.21437/JEP.2018-33

# 1 Pourquoi une mesure d'intelligibilité sous forme de décodage acoustico-phonétique ?

## 1.1 Intelligibilité et compréhensibilité de la parole

La perception de la parole est un processus complexe qui intègre à la fois un flux ascendant d'informations provenant du signal vocal mais aussi un flux descendant fondé sur les informations de haut niveau détenues par l'auditeur. Le flux ascendant (« bottom-up ») est principalement une opération de décodage acoustico-phonétique qui consiste à identifier les phonèmes à partir du signal de parole. Les phonèmes, pouvant être considérés comme les plus petites unités permettant d'opposer du sens, sont les éléments de base de l'intelligibilité du discours, c'est-à-dire du degré de précision avec lequel le message est compris par l'auditeur. Le décodage acoustico-phonétique est donc le processus fondamental pour mesurer perceptivement l'intelligibilité d'un locuteur.

Le flux descendant (« top-down ») fait appel chez l'auditeur à un ensemble d'informations qu'il détient à différents niveaux : la connaissance du lexique de façon générale, la connaissance du contexte de la situation de communication pouvant potentiellement restreindre considérablement le lexique de circonstance, la connaissance des communicants... De ce fait, lorsqu'un auditeur entend un énoncé dégradé, bruité ou phonétiquement appauvri, ces processus top-down entrent en jeu pour restaurer ce qui est distordu et optimiser l'intelligibilité du message (Warren et al., 1970). Les effets de lexicalité, c'est-à-dire le fait qu'une séquence sonore ou orthographique fasse référence à un mot de notre vocabulaire, sont notamment très forts. Les travaux de (Ganong, 1980) ont montré qu'en anglais, un son phonétiquement ambigu t/d sera préférentiellement perçu [d] s'il est placé devant une séquence [aʃ] en référence au mot « dash », et inversement, le même son sera perçu [t] devant une séquence [ask] en référence au mot « task ». Il faut remarquer qu'en français, le résultat serait inversé : un son phonétiquement ambigu t/d sera préférentiellement perçu [t] devant une séquence [aʃ] en référence au mot « tache » mais il sera perçu [d] s'il est placé devant une séquence [isk] en référence au mot « disque ». A ces effets de lexicalité s'ajoutent d'autres phénomènes comme la fréquence des mots (les mots usuels sont plus facilement reconnus), les règles phonotactiques de la langue (une séquence [vrsitʃ] est peu probable en français), le savoir partagé relatif au contexte de la conversation.

Dans le cas où nous nous intéressons à l'intelligibilité d'un locuteur produisant une parole atypique (production pathologique de la parole, apprentissage des langues, acquisition ou vieillissement), ces mécanismes top-down peuvent s'avérer gênants pour mesurer le degré de précision/perturbation dans la mesure où ils interviennent chez l'auditeur de façon variable et qu'ils peuvent, en conséquence, masquer des altérations présentes chez le locuteur. Le type de test choisi va plus ou moins donner de l'importance aux processus perceptifs descendants. Plus les mécanismes top-down sont impliqués chez l'auditeur, plus on s'écarte de l'évaluation de la performance du locuteur. Dans un cadre clinique, on s'éloigne de la mesure de l'altération en se plaçant sur le versant de l'invalidité, voire de son potentiel handicap au sens de la terminologie de l'OMS. C'est le cas des tests de compréhensibilité qui incluent du décodage acoustico-phonétique (processus ascendant inhérent à tous les tests), de l'accès lexical mais prennent également en compte le contexte de l'échange entre les interactants et tous les autres moyens que le patient met en œuvre pour se faire comprendre (gestes, mimiques, connaissances implicites…). C'est la raison pour laquelle la compréhensibilité reste difficilement quantifiable et qu'on préfère mesurer l'intelligibilité dont on peut obtenir des scores (Woisard et al., 2013).

### 1.2 Limiter les effets « top-down »

De façon classique, les tests d'intelligibilité sont effectués à partir de phrases ou de mots issus de listes de référence. Les limitations de ce type de test résident dans la capacité des auditeurs à restaurer les séquences distordues. Cet effet est d'autant plus fort que les auditeurs ont une connaissance forte des mots utilisés dans le test et que ces mots sont peu ambigus et donc fortement prédictibles. C'est généralement le cas des orthophonistes qui peuvent faire un usage si important de ces listes qu'ils/elles finissent par les connaître par cœur. On peut citer par exemple la BECD (Auzou et al., 2006) qui ne comporte que 50 mots. Le biais lié à cette connaissance et donc à la forte influence des mécanismes perceptifs descendants est un score d'intelligibilité surévalué car la restauration phonémique de l'auditeur rend « transparentes » les distorsions de production.

## 2 L'intelligibilité par le biais de 90 000 pseudo-mots

### 2.1 Construction et principes du test

La solution que nous proposons consiste à utiliser des pseudo-mots, c'est-à-dire des logatomes respectant les structures phonotactiques fréquentes du français, en grande quantité de façon à complètement neutraliser les effets de lexicalité ou d'apprentissage des items par les auditeurs. Au final, les auditeurs sont confrontés à une tâche de décodage acoustico-phonétique suivie d'une transcription écrite. Les détails de la construction du test sont donnés dans (Ghio et al., 2016). Le principe du test est de faire prononcer 52 pseudo-mots tirés aléatoirement d'une liste de 89346 formes possibles, sachant que chaque liste est, par construction, phonétiquement équilibrée. Les pseudo-mots ont été construits avec les formes $C(C)_1V_1C(C)_2V_2$ où $C(C)_i$ est une consonne isolée ou un groupe consonantique. Par exemples: stoumo, vurtant, muja, charou, leba, ranto…

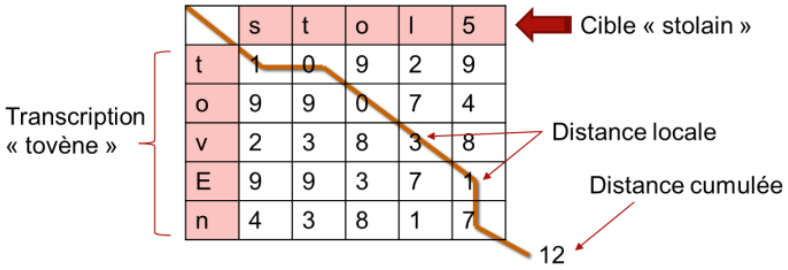
Pour permettre l'énonciation des pseudo-mots par les locuteurs, nous avons utilisé le logiciel PERCEVAL-LANCELOT (www.lpl-aix.fr/~lpldev/perceval/). Le locuteur est placé devant un écran sur lequel est affiché automatiquement le pseudomot à prononcer et une version sonore est produite de façon synchronisée. Cette double modalité, visuelle et auditive, permet de limiter les erreurs de lecture, les limitations auditives et attentionnelles. Etant donnée la taille importante du corpus (89346 formes possibles), les versions sonores sont issues de la synthèse Voxygen (voxygen.fr/). Le locuteur est alors enregistré. Ses enregistrements sont ensuite segmentés semi-automatiquement pour obtenir un fichier audio par logatome produit. L'ensemble des stimuli de tous les locuteurs est finalement soumis à un ensemble d'auditeurs dont la tâche est de transcrire ce qu'ils entendent via le logiciel LANCELOT.

### 2.2 Le traitement des transcriptions orthographiques

La consigne donnée aux auditeurs pour transcrire orthographiquement chaque pseudo-mot produit par les locuteurs est la suivante : « *Vous allez entendre des non-mots. Un non-mot est une combinaison de sons de la langue française qui n'a pas de signification (ex: gloutu). En respectant les règles de l'orthographe du français, vous devrez transcrire ce que vous entendrez. Certaines prononciations seront difficiles à identifier mais dans tous les cas, vous devrez proposer une transcription.* » Les auditeurs sont choisis natifs de langue française, sans problème auditif et ayant une bonne maitrise de l'orthographe.

Une fois les transcriptions orthographiques recueillies, l'objectif est d'en extraire une forme phonémique car le passage par l'orthographe n'est qu'une étape intermédiaire pour accéder à une représentation phonétique. Les transcriptions orthographiques sont donc phonétisées par l'algorithme LIA_PHON (Bechet, 2001) et elles sont comparées aux formes phonétiques attendues des pseudo-mots. Traditionnellement, par facilité de traitement, le résultat est binaire : correct ou incorrect. Pour dépasser cette évaluation sommaire, nous proposons un résultat analogique sous forme de distance à la cible.

Pour l'opération de comparaison, nous avons utilisé un algorithme de Wagner-Fischer qui intègre les phénomènes d'insertion, élision et substitution d'unités (Figure 1). Dans notre cas, ce calcul de distance de Levenshtein portant non pas sur des unités orthographiques mais sur les phonèmes, il nous est apparu important d'établir une distance locale entre unités (Ghio, 1997). En effet, sur les formes orthographiques, de façon traditionnelle, la distance entre 2 graphèmes est nulle s'ils sont égaux et vaut 1 s'ils sont différents. Dans le cas de phonèmes, il est possible d'apporter des nuances plus subtiles car, par exemple, on peut considérer qu'une confusion entre 2 voyelles n'a pas le même poids qu'entre une voyelle et une consonne sourde.




| | s | t | o | l | 5 |
|---|---|---|---|---|---|
| t | 1 | 0 | 9 | 2 | 9 |
| o | 9 | 9 | 0 | 7 | 4 |
| v | 2 | 3 | 8 | 3 | 8 |
| E | 9 | 9 | 3 | 7 | 1 |
| n | 4 | 3 | 8 | 1 | 7 |

*Figure 1 : Comparaison de 2 chaines phonémiques par l'algorithme de Wagner-Fischer (conventions : « 5 » = /ɛ̃/ , « E » = /ɛ/)*

# 3 L'établissement de la matrice de coût entre phonèmes

## 3.1 La métrique

La matrice de « coût » est un tableau qui contient le degré de dissimilitude entre phonèmes. Elle comporte les 35 phonèmes / a i u o ɔ e ɛ y œ ø ɔ̃ ã ɛ̃ œ̃ p t k b d g f s ʃ v z ʒ m n l R j w ɥ ñ ŋ / auxquels s'ajoutent divers archiphonèmes : Ô = /o/ ou /ɔ/, Ê = /e/ ou /ɛ/, Û = /ø/ ou /œ/, µ= /ɛ̃/ ou /œ̃/, & = /e/ ou /ɛ/ ou /ø/ ou /œ/. Pour le codage des unités phonologiques au format informatique, nous avons utilisé la convention de lexique.org (www.lexique.org/listes/liste_codes_phono.php) car elle a l'avantage de coder une unité sur un caractère, contrairement au codage SAMPA dont la correspondance se fait sur 1 ou 2 caractères, ce qui complique le codage.
Pour constituer la matrice, deux stratégies peuvent être adoptées:

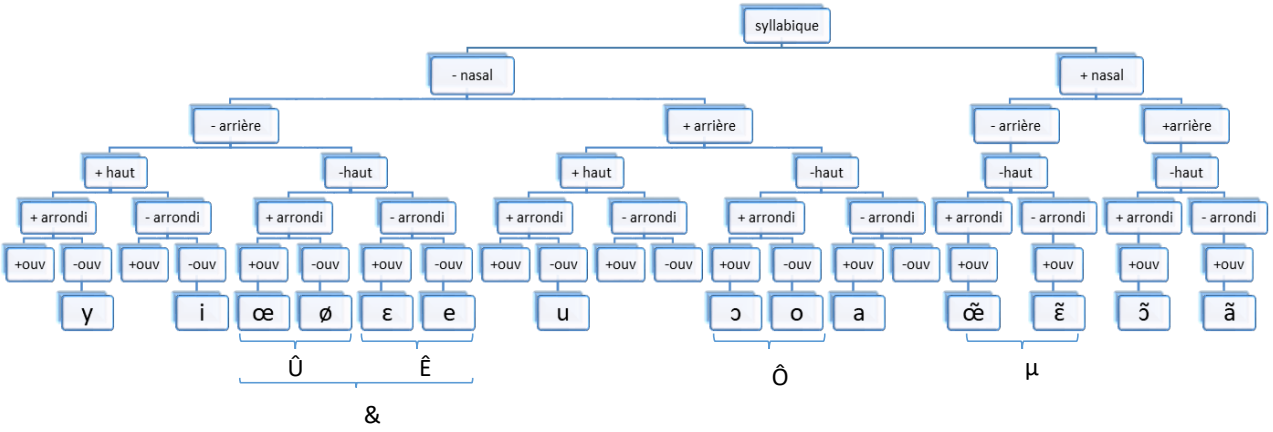
- Une mesure fondée sur les données. Dans ce cas-là, des procédures automatiques calculent statistiquement l'écart moyen entre phonèmes. Il s'agit alors de choisir un corpus représentatif ainsi qu'une métrique pertinente de comparaison.
- Une mesure fondée sur les connaissances. Dans ce cas-là, la distance entre phonèmes est attribuée a priori à partir de savoirs partagés.

Les résultats de nos tests d'intelligibilité pouvant être utilisés comme base d'apprentissage de mesures issus du traitement automatique, nous avons voulu éviter une forme de circularité et avons donc écarté la 1ère solution. Nous avons choisi la seconde méthode.

Afin de réduire son aspect arbitraire, nous avons fondé la comparaison sur la théorie des traits, c'est-à-dire sur le fait que les phonèmes peuvent être décomposés en un ensemble de traits qui les distingue (Jakobson et al., 1951). Il est facile de construire, à partir de cette décomposition, un espace multidimensionnel dans lequel chaque phonème est repéré géométriquement. La notion de traits imposant un caractère binaire (présent ou absent), les coordonnées des phonèmes dans l'espace multidimensionnel ne prennent que les valeurs 0 ou 1. Cela diminue grandement l'importance du choix de la norme. En effet, dans ce cas-là, la valeur donnée par une distance euclidienne $\left(d = \sqrt{\sum_i (x_i - y_i)^2}\right)$ est la racine carrée de celle fournie par une distance de norme 1 $\left(d = \sum_i |x_i - y_i|\right)$. Il n'existe qu'un effet de contraction que nous n'étudierons pas. Nous avons préféré utiliser la distance de norme 1, qui consiste finalement à compter le nombre de traits différents entre deux phonèmes.

## 3.2 La matrice de coût des voyelles

La Table 1 présente la décomposition en traits distinctifs des voyelles du français d'après Chomsky et Halle (1968). Nous avons remplacé la dénomination chomskyenne [+/- bas] par [+/- ouvert] car moins sujette à confusion avec le trait [+/- haut] qui n'est pas l'opposé du trait [+/- bas]. Dans ce cadre, les voyelles moyennes /e ø o/ sont [-haut ; -bas] et s'opposent respectivement à /ɛ œ ɔ/ qui sont [+bas], c'est-à-dire [+ouvert] dans notre dénomination. La décomposition en arbre permet de mettre en évidence la notion d'archiphonème, c'est-à-dire la sous-spécification d'un trait. Ainsi, les archiphonèmes Ê={e, ɛ}, Û={œ, ø}, Ô={o, ɔ} sont des unités où le trait d'ouverture n'est pas spécifié ; de même, µ={ œ̃, ɛ̃ } et &={ Ê, Û} neutralisent le trait d'arrondissement (labialisation).




| | a | i | u | o | e | y | ø | ɛ | ɔ | œ | Ô | Û | Ê | & | ã | ɛ̃ | ɔ̃ | œ̃ | µ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| nasal | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| arrière | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| haut | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| arrondi | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | | 0 | 0 | 1 | 1 | |
| ouvert | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | | | | | 1 | 1 | 1 | 1 | 1 |

*Table 1 : Décomposition en traits des voyelles du français sous forme d'arbre et de matrice*

Cette décomposition permet ainsi de dresser une matrice de distances entre voyelles par comptage du nombre de traits différents entre chaque phonème (Table 2). La matrice est symétrique.

| | a | i | u | o | e | y | ø | ɛ | ɔ | œ | ã | ɛ̃ | ɔ̃ | œ̃ | Ê | Ô | Û | µ | & |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| a | 0 | 3 | 3 | 2 | 2 | 4 | 3 | 1 | 1 | 2 | 1 | 2 | 2 | 3 | 1 | 1 | 2 | 2 | 1 |
| i | 3 | 0 | 2 | 3 | 1 | 1 | 2 | 2 | 4 | 3 | 4 | 3 | 5 | 4 | 1 | 3 | 2 | 3 | 1 |
| u | 3 | 2 | 0 | 1 | 3 | 1 | 2 | 4 | 2 | 3 | 4 | 5 | 3 | 4 | 3 | 1 | 2 | 4 | 2 |
| o | 2 | 3 | 1 | 0 | 2 | 2 | 1 | 3 | 1 | 2 | 3 | 4 | 2 | 3 | 2 | 0 | 1 | 3 | 1 |
| e | 2 | 1 | 3 | 2 | 0 | 2 | 1 | 1 | 3 | 2 | 3 | 2 | 4 | 3 | 0 | 2 | 1 | 2 | 0 |
| y | 4 | 1 | 1 | 2 | 2 | 0 | 1 | 3 | 3 | 2 | 5 | 4 | 4 | 3 | 2 | 2 | 1 | 3 | 1 |
| ø | 3 | 2 | 2 | 1 | 1 | 1 | 0 | 2 | 2 | 1 | 4 | 3 | 3 | 2 | 1 | 1 | 0 | 2 | 0 |
| ɛ | 1 | 2 | 4 | 3 | 1 | 3 | 2 | 0 | 2 | 1 | 2 | 1 | 3 | 2 | 0 | 2 | 1 | 1 | 0 |
| ɔ | 1 | 4 | 2 | 1 | 3 | 3 | 2 | 2 | 0 | 1 | 2 | 3 | 1 | 2 | 2 | 0 | 1 | 2 | 1 |
| œ | 2 | 3 | 3 | 2 | 2 | 2 | 1 | 1 | 1 | 0 | 3 | 2 | 2 | 1 | 1 | 1 | 0 | 1 | 0 |
| ã | 1 | 4 | 4 | 3 | 3 | 5 | 4 | 2 | 2 | 3 | 0 | 1 | 1 | 2 | 2 | 2 | 3 | 1 | 2 |
| ɛ̃ | 2 | 3 | 5 | 4 | 2 | 4 | 3 | 1 | 3 | 2 | 1 | 0 | 2 | 1 | 1 | 3 | 2 | 0 | 1 |
| ɔ̃ | 2 | 5 | 3 | 2 | 4 | 4 | 3 | 3 | 1 | 2 | 1 | 2 | 0 | 1 | 3 | 1 | 2 | 1 | 2 |
| œ̃ | 3 | 4 | 4 | 3 | 3 | 3 | 2 | 2 | 2 | 1 | 2 | 1 | 1 | 0 | 2 | 2 | 1 | 0 | 1 |
| Ê | 1 | 1 | 3 | 2 | 0 | 2 | 1 | 0 | 2 | 1 | 2 | 1 | 3 | 2 | 0 | 2 | 1 | 1 | 0 |
| Ô | 1 | 3 | 1 | 0 | 2 | 2 | 1 | 2 | 0 | 1 | 2 | 3 | 1 | 2 | 2 | 0 | 1 | 2 | 1 |
| Û | 2 | 2 | 2 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 3 | 2 | 2 | 1 | 1 | 1 | 0 | 1 | 0 |
| µ | 2 | 3 | 4 | 3 | 2 | 3 | 2 | 1 | 2 | 1 | 1 | 0 | 1 | 0 | 1 | 2 | 1 | 0 | 1 |
| & | 1 | 1 | 2 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 2 | 1 | 2 | 1 | 0 | 1 | 0 | 1 | 0 |

*Table 2 : matrice de coût des voyelles (⇔ nombre de traits différents entre les voyelles)*

## 3.3 La matrice de coût des consonnes

Dans la décomposition des consonnes du français, un certain nombre de traits est clairement défini :

- Le trait vocalique (+/- sonant) distingue les obstruantes (occlusives et fricatives : -sonant) des consonnes liquides (l R), nasales (m n ñ) et semi-voyelles (j w ɥ ) : +sonant
- Le trait de nasalité distingue les consonnes nasales (+nasal) des orales (-nasal)
- Le trait de voisement distingue les consonnes sonores (voisées) des sourdes (-vois)
- Le trait de continuité distingue les occlusives (-cont) des fricatives (+cont).
  Parmi les consonnes vocaliques, Chomsky et Halle (1968) précisent p.317 que les occlusives nasales sont considérées comme interrompues (-cont). Les auteurs précisent enfin que le cas de /l/ et /r/ est complexe mais finissent par proposer un trait (+cont) à /r/ et (-cont) à /l/. Cette caractérisation est confirmée dans Clements (2005) p.47.

En revanche, les traits relatifs au lieu d'articulation de la consonne posent de multiples problèmes. En effet, d'après l'alphabet phonétique international (www.internationalphoneticalphabet.org), les consonnes du français sont articulées selon 7 lieux différents qui peuvent être regroupés en 3 grandes classes d'articulation : les labiales, les dentales et les vélo-palatales (Table 3).

| | Bilabial | Labio-dental | Dental | Alveolar | Post-alveolar | Palatal | Velar |
|---|---|---|---|---|---|---|---|
| Plosive | p b | | | t d | | | k g |
| Nasal | m | | | n | | ɲ | ŋ |
| Fricative | | f v | | s z | ʃ ʒ | | |
| approximant | | | | l | | j | |
| | Labiales | | Dentales | | Vélo-Palatales | | |

*Table 3 : lieu d'articulation des consonnes du français (d'après l'IPA)*

Dans une approche totalement phonologique, Chomsky et Halle (1968) proposent p.223 la décomposition selon les deux traits +/- coronal (pointe de la langue) et +/- antérieur, ce qui donne la Table 4 ci-dessous. Cette décomposition place alors /p/ (-cor, +ant) à un trait de distance de /t/ (+cor) et à un trait de distance de /k/ (-ant). En revanche, il place /t/ (+cor, +ant) à deux traits d'écart de /k/ (-cor, -ant), ce qui n'est pas très satisfaisant d'un point de vue articulatoire où il semblerait logique de respecter l'ordre /p t k/, c'est-à-dire /t/ équidistant de /p/ et /k/, /p/ et /k/ étant plus éloignés.

| | + coronal | -coronal |
|---|---|---|
| +antérieur | Dental : t d s z | Labial : p b f v |
| -antérieur | Palato-alveolaire : rien en français | Vélaire : k g (ʃ ʒ) |

*Table 4 : traits relatifs au lieu d'articulation d'après Chomsky et Halle (1968)*

Clements (2005) propose une décomposition en 3 traits exclusifs : labial, coronal et dorsal qui reflètent directement les 3 lieux décrits en Table 3. Nous estimons qu'il y a là une sur spécification car 2 traits seulement sont nécessaires pour coder 3 états. Nous avons finalement opté pour les travaux de Jakobson et al. (1951) qui proposent 2 traits acoustiques permettant une distinction adéquate :

- Le trait compact/diffus : "the consonants articulated against the hard or soft palate (velars and palatals) are more compact than the consonants articulated in the front part of the mouth." (Jakobson et al., 1951, p.27)
- Le trait grave/aigu : " gravity characterizes labial consonants as against dentals, as well as velars vs. palatals" (Jakobson et al., 1951, p.30)

Nous obtenons finalement la décomposition des consonnes en traits (Table 5) et la matrice de distances entre consonnes (Table 6).

| | p | t | k | b | d | g | f | s | S | v | z | Z | m | n | N | l | R | j | w | ɥ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| vocalique | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| continu | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| nasal | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| voisé | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| compact | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| aigu | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |

*Table 5 : Décomposition en traits des consonnes du français*

| | p | t | k | b | d | g | f | s | S | v | z | Z | m | n | N | l | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p | 0 | 1 | 2 | 1 | 2 | 3 | 1 | 2 | 3 | 2 | 3 | 4 | 3 | 4 | 5 | 3 | 4 |
| t | 1 | 0 | 1 | 2 | 1 | 2 | 2 | 1 | 2 | 3 | 2 | 3 | 4 | 3 | 4 | 2 | 3 |
| k | 2 | 1 | 0 | 3 | 2 | 1 | 3 | 2 | 1 | 4 | 3 | 2 | 5 | 4 | 3 | 3 | 4 |
| b | 1 | 2 | 3 | 0 | 1 | 2 | 2 | 3 | 4 | 1 | 2 | 3 | 2 | 3 | 4 | 2 | 3 |
| d | 2 | 1 | 2 | 1 | 0 | 1 | 3 | 2 | 3 | 2 | 1 | 2 | 3 | 2 | 3 | 1 | 2 |
| g | 3 | 2 | 1 | 2 | 1 | 0 | 4 | 3 | 2 | 3 | 2 | 1 | 4 | 3 | 2 | 2 | 3 |
| f | 1 | 2 | 3 | 2 | 3 | 4 | 0 | 1 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 4 | 3 |
| s | 2 | 1 | 2 | 3 | 2 | 3 | 1 | 0 | 1 | 2 | 1 | 2 | 5 | 4 | 5 | 3 | 2 |
| S | 3 | 2 | 1 | 4 | 3 | 2 | 2 | 1 | 0 | 3 | 2 | 1 | 6 | 5 | 4 | 4 | 3 |
| v | 2 | 3 | 4 | 1 | 2 | 3 | 1 | 2 | 3 | 0 | 1 | 2 | 3 | 4 | 5 | 3 | 2 |
| z | 3 | 2 | 3 | 2 | 1 | 2 | 2 | 1 | 2 | 1 | 0 | 1 | 4 | 3 | 4 | 2 | 1 |
| Z | 4 | 3 | 2 | 3 | 2 | 1 | 3 | 2 | 1 | 2 | 1 | 0 | 5 | 4 | 3 | 3 | 2 |
| m | 3 | 4 | 5 | 2 | 3 | 4 | 4 | 5 | 6 | 3 | 4 | 5 | 0 | 1 | 2 | 2 | 3 |
| n | 4 | 3 | 4 | 3 | 2 | 3 | 5 | 4 | 5 | 4 | 3 | 4 | 1 | 0 | 1 | 1 | 2 |
| N | 5 | 4 | 3 | 4 | 3 | 2 | 6 | 5 | 4 | 5 | 4 | 3 | 2 | 1 | 0 | 2 | 3 |
| l | 3 | 2 | 3 | 2 | 1 | 2 | 4 | 3 | 4 | 3 | 2 | 3 | 2 | 1 | 2 | 0 | 1 |
| R | 4 | 3 | 4 | 3 | 2 | 3 | 3 | 2 | 3 | 2 | 1 | 2 | 3 | 2 | 3 | 1 | 0 |

*Table 6 : matrice de coût des consonnes (⇔ nombre de traits différents entre les consonnes)*

## 3.4 Les distances inter macro-classes

Les semi-consonnes /j w ɥ/ ont été placées de façon identique à leur équivalent /i u y/ mais avec le trait de syllabicité en moins (-syll). En effet, ces phonèmes ne peuvent à eux seuls constituer une syllabe (Chomsky & Halle, 1968). Dans leurs distances aux consonnes, elles ont été décomposées comme présentées en Table 5.

Par rapport aux voyelles, les consonnes ont été placées à une distance supérieure à la distance maximale entre voyelles (d=6). En tenant compte de la classification de Dell (1985),

| | | | | |
|---|---|---|---|---|
| non syllabique | consonantique | non vocalique | sourd | Obstruantes sourdes |
| | | | sonore | Obstruantes sonores |
| | non consonantique | vocalique | | consonnes nasales et liquides |
| syllabique | | | | semi-voyelles |
| | | | | voyelles |

nous avons ensuite respecté la hiérarchie suivante :

Voyelles < Liquides < Nasales < Obstruantes sonores < Sourdes

Au final, nous obtenons une matrice de « coût » qui contient le degré de dissimilitude entre les 35 phonèmes retenus pour le français.

# 4 Application : La mesure d'intelligibilité de patients avec traitement du cancer des voies aériennes supérieures

Le protocole décrit précédemment a été utilisé dans le cadre du projet C2SI (Carcinologic Speech Severity Index) dont l'objectif est d'obtenir une mesure de l'impact des traitements des cancers de la cavité buccale et du pharynx sur la production de la parole par l'Indice de sévérité des troubles de la production de la parole à la fois par des méthodes perceptives et par traitement automatique de la parole.

126 locuteurs (41 sujets sains et 85 patients) enregistrés dans le service d'oncoréhabilation de l'Oncopole à Toulouse ont produit chacun 52 pseudo-mots tirés aléatoirement de la liste de 89346 formes possibles. 40 auditeurs ont retranscrit ces productions, chaque pseudo-mot d'un locuteur étant transcrit par 3 auditeurs différents. Ces tests se sont déroulés au sein du Centre d'Expérimentation sur la Parole (www.lpl-aix.fr/~cep) du Laboratoire Parole et Langage à Aix-en-Provence. Les transcriptions orthographiques ont été phonétisées et comparées aux formes phonétiques attendues des pseudo-mots par l'algorithme décrit au §2.2

De façon globale, les résultats montrent que les formes perçues chez les sujets sains sont en moyenne à une distance de 0.48 trait/phonème (sdev= 0.22) par rapport aux formes attendues alors que cette distance passe à 1.28 (sdev=0.63) pour les patients. En effectuant une transformation logarithmique du score, les distributions deviennent normales (test de Shapiro ; $p > 0.05$) et les variances homogènes (test de Bartlett ; $p > 0.05$). La différence entre les deux groupes est significative ($p < 0.01$).

Dans l'avenir, nous allons nous employer à vérifier l'équivalence des listes, c'est-à-dire que nous allons mesurer quels sont les écarts obtenus sur un même locuteur produisant plusieurs listes. De plus, ces mesures vont être comparées à des évaluations cliniques globales ainsi qu'à des mesures acoustiques automatiques (Astésano et al. , 2018).

# 5 Conclusion

Nous avons mis au point un test d'intelligibilité à partir d'une importante cohorte de pseudo-mots répondant aux contraintes phonotactiques du français. Cette méthode a été testée sur 126 locuteurs en milieu hospitalier sans obstacle majeur. La transcription orthographique par des auditeurs est suivie d'une transformation graphème-phonème puis d'une comparaison sophistiquée à la cible phonétique attendue. Cette comparaison fondée sur un calcul de traits distinctifs a l'immense avantage de créer une métrique progressive différente des approches traditionnelles qui se contentent de compter le nombre d'occurrences correctes. La construction même du test ne permet aucune restauration phonémique par les auditeurs des séquences mal produites par les locuteurs. Il n'y a donc pas d'effet plafond, ce qui pourra permettre de quantifier finement par exemple des effets thérapeutiques. Pour conclure, le test semble discriminant en ce qui concerne la mesure de la performance articulatoire des locuteurs.

# Remerciements

Ce travail fait partie du projet C2SI (Carcinologic Speech Severity Index) financé par l'Institut National du Cancer dans le cadre de projets libres de recherche en Sciences Humaines et Sociales, Epidémiologie et Santé Publique. L'investigatrice principale est Virginie Woisard du CHU Larrey à Toulouse. Nous remercions la Sté Voxygen pour avoir synthétisé les 89346 stimuli du corpus. Nous remercions le personnel du CEP (www.lpl-aix.fr/~cep), en particulier Carine André, pour la réalisation des tests de perception.

# Références


ASTESANO C. , BALAGUER M., FARINAS J., FREDOUILLE C., GAILLARD P., GHIO A., GIUSTI L. et al. (2018), Carcinologic Speech Severity Index Project: A Database of Speech Disorders Productions to Assess Quality of Life Related to Speech After Cancer, LREC, 7-12 May 2018, Miyazaki (Japan)

AUZOU P, ROLLAND-MONNOURY V. (2006), Batterie d'évaluation de la dysarthrie, *1st ed. Isbergues: Ortho Edition*.

BECHET F (2001), LIA_PHON : UN SYSTEME COMPLET DE PHONETISATION DE TEXTES, TRAITEMENT AUTOMATIQUE DES LANGUES - TAL - VOLUME 42 NUMERO 1 - PP 47-67, 2001

CLEMENTS G.N. (2005), The role of features in speech sound inventories In Raimy & Cairns, eds., *Contemporary Views on Architecture and Representations in Phonological Theory*. Cambridge, MA: MIT Press, p 19-68

CHOMSKY N., HALLE M. (1968), *The Sound Pattern of English*. New York: Harper & Row.

DELL F. (1985), LES REGLES ET LES SONS, HERMANN, PARIS.

GANONG W. (1980). Phonetic categorization in auditory word perception. Journal of Experimental Psychology. *Human perception and performance*, 6 (1), 110-125.

GHIO A., GIUSTI L., BLANC E., PINTO S., LALAIN M, ROBERT D., FREDOUILLE C., WOISARD V. (2016) Quels tests d'intelligibilité pour évaluer les troubles de production de la parole ?. *Journées d'Etude sur la Parole*, Paris, France, p.589-596

GHIO A.(1997) Achile : un dispositif de décodage acoustico-phonétique et d'identification lexicale indépendant du locuteur à partir de modules mixtes. Thèse de l'Université d'Aix Marseille, 1997.

JAKOBSON R., FANT G., HALLE M. (1951), "Preliminaries to speech analysis", MIT Press, Cambridge.

WARREN RM., WARREN RP. (1970), Auditory illusions and confusions. *Sci. Am*.; 223, 30-36

WOISARD V., ESPESSER R., GHIO A., DUEZ D. (2013). De l'intelligibilité à la compréhensibilité de la parole, quelles mesures en pratique clinique ? Revue de laryngologie, otologie, rhinologie, vol. 1, no. 134. 2013, p. 27-33.

# Variabilité du geste linguo-palatal. Le cas du russe

Ekaterina Biteeva Lecocq  Nathalie Vallée  Silvain Gerber  Christophe Savariaux
GIPSA-Lab, UMR 5216, CNRS & Université Grenoble Alpes, BP25 38040 Grenoble cedex 9, France
ekaterina.biteeva@gipsa-lab.fr, nathalie.vallee@gipsa-lab.fr,
silvain.gerber@gipsa-lab.fr, christophe.savariaux@gipsa-lab.fr

## RÉSUMÉ

Le trait *palatal*, souvent discuté, ne capture pas toute la complexité et la diversité des réalisations consonantiques rangées sous ce trait. Le russe a développé des consonnes palatales et palatalisées permettant ainsi d'observer les réalisations de ce trait au sein du même système. À partir d'une analyse de données acquises auprès de 9 locuteurs grâce à un articulographe électromagnétique, nous avons examiné les patterns spatiaux et temporels des configurations linguales dans la réalisation des consonnes /tʲ ʧ ʃ ʃː/ du russe. Leur variation est observée selon les facteurs locuteur, accent et position de la syllabe. Les résultats confirment en partie ceux de Biteeva Lecocq et *al*. (2016). Les analyses montrent une importante variabilité du geste linguo-palatal en fonction des locuteurs. En revanche, aucun effet des deux autres facteurs testés n'a été observé. Ces deux points sont discutés à la lumière de travaux antérieurs.

## ABSTRACT

**Palatal gesture variability in Russian: speaker, stress and syllabic structure effects.**
The *palatal* feature is often discussed because it does not capture all the complexity of a palatal articulation. Moreover, it does not take into account the great range of consonantal realizations associated with this phonological feature. Russian has developed palatal and palatalized consonants and hence makes possible the observation of different realizations of the *palatal* feature within a phonological system. From a set of articulatory data acquired using an electromagnetic articulograph, we analyzed the variation of spatial and temporal patterns of the tongue configurations during the production of Russian consonants / tʲ ʧ ʃ ʃː/ according to the following factors: speaker, stress and position in the syllable. The results confirm partially those of Biteeva Lecocq et *al*. (2016). The analyzes show a large interspeaker variability and no effect of the two other tested factors. These two points are discussed considering previous studies.

MOTS-CLÉS : Palatalisation, geste palatal, variabilité, contrôle articulatoire, russe, EMA.

KEYWORDS: Palatalization, palatal gesture, variability, articulatory control, Russian, EMA.

# 1 Introduction

Les linguistes se sont souvent interrogés à propos de la nature des consonnes [+ palatal] en russe et dans d'autres langues. Les aspects articulatoires ont été abordés dans les différents travaux de Skalozub (1963) et de Kuznetsova (1969), de Straka (1965) sur la force articulatoire des palatales et palatalisées, ou encore plus récemment dans l'étude IRM de Kedrova et *al*. (2008). Les caractéristiques acoustiques ont été abordées dans différents travaux de Kochetov (2002). Une étude de Kavitskaya (2002) a été consacrée à la perception des consonnes palatalisées. Par ailleurs, les aspects phonologiques de ce type consonantique ont été étudiés en lien avec la phonétique par Keating (1988, 1991, 1993), Recasens (1990), Recasens et *al*. (1993, 1995) et Recasens et Romero (1997). Ces travaux ont montré entre autres que la question du contrôle du geste linguo-palatal


10.21437/JEP.2018-34

suscite encore des discussions en phonétique et en phonologie. Les phonéticiens sont en désaccord quant à la caractérisation du geste lingual entre consonnes palatales et palatalisées alors que les caractéristiques phonétiques sont beaucoup plus stables pour les consonnes labiales, dentales ou vélaires. Selon Recasens et Romero (1997), les palatales sont des articulations simples, au contraire des palatalisées réalisées à partir d'une superposition de deux articulations, primaire et secondaire. Selon Keating (1988, 1993), les palatales et palatalisées sont des segments complexes qui impliquent dans leur production deux gestes articulatoires (partie antérieure et dos de la langue) et donc, deux contrôles moteurs différents. D'autre part, le trait *palatal* ne capture pas toute la complexité d'une articulation palatale et surtout ne prend pas en compte la diversité des réalisations consonantiques rangées sous ce trait phonologique (Sagey, 1986). Par exemple, la qualification de consonne *molle* dans la littérature est utilisée pour désigner aussi bien les consonnes palatales que les palatalisées, alors que certains des travaux suggèrent des différences articulatoires et acoustiques entre ces deux types consonantiques (Straka, 1965 ; Ladefoged et Maddieson, 1996).

Le russe et d'autres langues slaves ont fréquemment développé dans leurs systèmes phonologiques des consonnes palatalisées, appelées dans la tradition russophone consonnes *molles* ou *mouillées* par opposition aux consonnes dites *dures*, alors que ce phénomène est peu rencontré dans d'autres langues (Maddieson, 1984 ; Vallée et *al*., 1999). En russe la palatalisation connait un fort rendement phonologique car la plupart des consonnes possèdent un équivalent palatalisé. Il est assez facile d'y trouver des paires minimales qui ne se distinguent que par le trait *palatal* : /bil/~/bilʲ/ être *(1re, 2e, 3e pers., sg., passé, imperfect.)* ~ événement *(nom., acc., sg.)*, /mat/~/mʲat/ tapis de sport *(nom., acc., sg.)* ~ froissé (*adj., forme courte*), /nos/~/nʲos/ nez *(nom., acc., sg.)* ~ porter *(1re, 2e, 3e pers., sg., passé, imperfect.)*. D'autre part, en russe la palatalisation peut être le résultat d'une assimilation régressive : /kostʲ/ > [kosʲtʲ] os *(nom., acc., sg.*), /ˈsdʲe.latʲ/ > [zʲdʲelətʲ] faire *(inf., perfect.)*. Le russe qui posséde également six consonnes palatales /ʧ ʃ ʃʲ: ʒ ʒʲ: j/ offre ainsi une gamme étendue de réalisations du trait *palatal*, qui plus est à l'intérieur du même système.

Nous avons choisi de nous intéresser à la nature articulatoire du phénomène de palatalisation. Le niveau articulatoire est d'autant plus complexe qu'il nécessite de prendre en compte et d'observer des stratégies individuelles. Notre étude expérimentale propose de caractériser le geste linguo-palatal impliqué dans la réalisation des différents types de consonnes palatalisées et palatales du russe et d'observer la variabilité de ce geste en fonction des facteurs locuteur, position de l'accent et type de structure syllabique. Notre objectif est d'examiner en fonction de ces différents facteurs extralinguistiques et linguistiques, le pattern spatial lors de l'atteinte de la cible articulatoire de la consonne ainsi que l'organisation temporelle de 4 points de référence sur la langue (pattern temporel) dans la phase d'atteinte de cette cible. À notre connaissance, peu d'études antérieures se sont penchées sur la variabilité des réalisations, hormis la variabilité liée aux différences de contexte vocalique (Recasens et *al*., 1993). Ainsi, l'étude du timing nous permettra d'examiner la question de la superposition d'un geste palatal à une articulation primaire lors de la production de consonnes palatalisées et de caractériser le contrôle du geste palatal, à savoir si le contrôle de la langue est global ou plutôt différencié selon les zones linguales, et s'il est différent entre palatalisées et palatales.

# 2 Méthodologie

## 2.1 Hypothèses

Dans le cadre de notre étude expérimentale et après un examen des travaux antérieurs nous avons mis à l'épreuve les hypothèses suivantes :

1. Si l'articulation palatale est secondaire, l'élévation de l'apex précède le geste palatal ; si l'articulation palatale est primaire, le dos de la langue s'élève en premier (Lindblom, Maddieson, 1988 ; Recasens, Romero, 1997) ;
2. Si l'accent impacte le geste lingual, on s'attend à ce qu'il soit plus ample et plus uniforme (Straka, 1963 ; Kelso et *al*., 1986 ; Fougeron, 1998) ;

3. Si la position dans la syllabe impacte le geste lingual, on s'attend à ce que le timing des différentes parties de la langue soit plus synchrone en coda en raison d'une réduction de l'amplitude du mouvement articulatoire dans cette position (Browman, Goldstein, 1988, 1995 ; Byrd, 1995 ; Kingston, 2008).

## 2.2 Participants et stimuli

Neuf locuteurs de langue maternelle russe, âgés de 22 à 42 ans ont participé à l'étude. Le corpus constitué comporte 40 mots contenant des consonnes palatalisées et palatales en position accentuée *vs* atone et en attaque syllabique *vs* coda. Dans cet article nous traitons les résultats pour les consonnes coronales /t tʲ ʧ ʃ ʃʲ:/ (table 1) qui présentent un fort rendement en russe pour l'opposition [±palatal]. Les mots contenant la consonne cible ont été insérés dans une phrase porteuse /ti ˈvʲi.dʲe.la/ *cible* /dva (trʲi) ˈra.za/ 'tu as vu '*cible*' deux (trois) fois' pour faciliter le repérage des frontières lors de la segmentation et neutraliser les effets suprasegmentaux. Les énoncés ont été présentés dans un ordre aléatoire en série de 10 éléments 5 fois pour chaque locuteur. La consigne donnée était de répéter à voix haute à un débit normal les énoncés présentés sur un écran.

| Paire minimale | Tonique *vs* atone | Attaque *vs* coda |
|---|---|---|
| /tuk/~/tʲuk/ bruit provoqué par un coup sur une surface dure *(onomat.)* ~ baluchon *(nom., acc., sg)* | /ˈtʲe.lo/ *vs* /tʲe.ˈla/ corps *(nom., acc., sg.) vs* corps *(nom., acc., pl.)*<br>/ˈʧer.tʲi/~/ʧer.ˈtʲi/ diablotin *(nom., pl.)* ~ tracer *(2e pers., sg., imperfect., injonc.)*<br>/ʃʲ:it/ *vs* /ʃʲ:i.ˈta/ bouclier *(nom., acc., sg.) vs* bouclier *(gén., sg.)*<br>/ʃestʲ/ *vs* /ʃes.ˈtʲi/ six *(nom., acc.) vs* six *(gén., dat., loc.)* | /tʲap/ *vs* /matʲ/ faire qqch à la va-vite *(onomat.) vs* mère *(nom., acc., sg.)*<br>/ʧem/ *vs* /mʲeʧ/ que *(conjonc. de subord.) vs* épée *(nom., acc., sg.)*<br>/ʃʲ:elʲ/ *vs* /lʲeʃʲ:/ fente *(nom., acc., sg.) vs* brème *(nom., sg.)*<br>/ʃov/ *vs* /voʃ/ suture *(nom., acc., sg.) vs* poux *(nom., acc., sg.)* |

TABLE 1 : Contrastes observés.

## 2.3 Protocole

Les données ont été acquises dans la chambre anéchoïde du GIPSA-Lab à l'aide d'un articulographe électromagnétique 2D (AG200 Carstens®), fréquence d'échantillonnage 200 Hz. Quatre bobines étaient collées sur la langue : au niveau de la pointe et de la lame, et au niveau du dos de la langue (*mid-* et *post-dorsum*, *infra*). Deux bobines de référence étaient collées au niveau de la racine du nez et des incisives supérieures pour la correction des mouvements de la tête et une bobine collée sur les incisives inférieures pour capturer le mouvement mandibulaire. Les données acoustiques ont été enregistrées avec un Marantz PMD 670 (échantillonnage 22,05 kHz), micro AKG C1000S. Les locuteurs étaient assis face à un écran 17 pouces AG NEOVO X-17A où s'affichaient les énoncés. A la fin de l'enregistrement le tracé du palais était réalisé par l'expérimentateur à l'aide d'une bobine collée sur son index et déplacée du fond de la cavité buccale vers les incisives supérieures. Toutefois, aucune détermination du lieu d'articulation ne peut être faite en fonction du palais acquis car s'assurer de l'immobilité de la tête du locuteur pendant la séance d'enregistrement est impossible.

## 2.4 Mesures et analyses

Les trajectoires des articulateurs ont été analysées dans l'environnement Matlab avec le logiciel interne TRAP. Les événements minimum et maximum des courbes de déplacement ont été repérés automatiquement à partir des passages par zéro de la courbe de vitesse de chacun des articulateurs. Les cibles consonantiques /t tʲ ʧ ʃ ʃʲ:/ sont considérées comme atteintes lorsque la pointe de la langue atteint sa position maximale.

Dans un premier temps, nous avons observé et analysé la configuration spatiale du geste lingual en choisissant comme caractérisation la position en X (degré d'avancement de la langue dans la cavité

buccale) et en Y (hauteur de la langue) des trois localisations lame, mid- et post-dorsum lorsque la pointe de la langue était à sa position maximale, instant correspondant à l'atteinte de la cible consonantique. Les données ont été recalées par rapport au plan occlusal utilisé comme axe des abscisses dans les représentations.

Dans un second temps nous avons analysé les patterns temporels, c'est-à-dire l'organisation temporelle de quatre points de référence sur la langue (apex, lame, mid- et post-dorsum). Pour chacun des quatre points l'écart (Δt) entre le temps de référence et le temps où chaque point atteint son maximum a été mesuré afin de déterminer le timing des maxima atteints par les quatre points de la langue en fonction de leur position dans la syllabe tonique ou atone et selon leur position en attaque ou en coda. Nous avons choisi comme geste de référence du début du mouvement lingual pour la réalisation de la consonne, le point le plus bas de la mandibule affecté à la réalisation de la voyelle qui précède la consonne étudiée. Ce point est l'instant à partir duquel la mandibule qui est le support de l'articulateur langue remonte pour réaliser l'articulation consonantique. Il nous semble que cet événement temporel est le plus à même de rendre compte du début du geste lingual de la production consonantique.

Des analyses statistiques ont été effectuées sur la variable réponse *durée* (ms) et l'influence sur celle-ci des facteurs tels que consonne (t, tʲ, ʧ, ʃ et ʃː), accent (syllabe accentuée *vs* atone) et position dans la syllabe (attaque *vs* coda). Pour examiner un éventuel impact de la force articulatoire nous avons créé une variable *renforcement articulatoire* qui se décline en trois modalités : attaque de syllabe tonique (Accent*Onset), coda de syllabe tonique (Accent*Coda) et attaque de syllabe atone (Atone*Onset). Nous avons également choisi de réaliser deux analyses distinctes, l'une uniquement avec les consonnes /t/ et /tʲ/ dans la modalité Accent*Onset, l'autre sans la consonne /t/, car aucune mesure pour celle-ci n'a été réalisée en dehors de ladite modalité. Le choix d'utiliser un modèle mixte a été fait avec *Locuteur* comme effet aléatoire et *Accent*, *Position dans la syllabe* comme effet fixe. Le modèle linéaire mixte a été réalisé à l'aide de la fonction lme du package nlme du logiciel R. D'autre part, les mesures ont été prises simultanément sur les quatre localisations de la langue et sont de ce fait corrélées entre elles ce qui nécessite la modélisation (à l'aide des paramètres weights et correlation de la fonction lme) de la matrice de variance-covariance du modèle (Bazzoli et *al*., 2016). La vérification des conditions d'applications des modèles a été effectuée en réalisant un diagnostic graphique des résidus. A partir du modèle nous avons effectué des comparaisons multiples (fonction glht du package multcomp de R) pour (1) analyser l'organisation temporelle des quatre bobines (localisations sur la langue) lors de la production d'une consonne donnée, (2) pour comparer les différences dans l'organisation temporelle des quatre bobines entre les types de consonnes et (3) pour comparer les positions des quatre bobines pour une consonne donnée en fonction des modalités du facteur *renforcement articulatoire*.

# 3 Résultats

## 3.1 Patterns spatiaux

Les figures qui suivent présentent les positions des 4 points de référence apex, lame, mid- et post-dorsum par locuteur et par consonne. Par manque de place nous présentons dans cette section uniquement les résultats pour /t/ et /tʲ/ dans la paire minimale /tuk/~/tʲuk/ et pour /ʃ/ et /ʃː/ dans les mots /ʃov/ *vs* /ʃː elʲ/. Nous décrivons ci-après les stratégies articulatoires que les locuteurs emploient pour réaliser des consonnes [± palatal] de leur langue.

**/tuk/~/tʲuk/**

Pour 7 locuteurs de notre étude /t/ est apicale : la pointe de la langue atteint son maximum dans la zone dento-alvéolaire. En revanche pour 3 de ces locuteurs, la configuration des points lame, mid- et post-dorsum est quasi alignée dans la cavité buccale alors que pour les 3 autres la forme linguale est plus ou moins concave au niveau de ces localisations. Enfin pour la septième participante, /t/ est apicale avec toutefois les positions des bobines mid- et post-dorsum plus élevées que la partie antérieure de la langue. Pour 2 des 9 locuteurs l'occlusion est apico-laminale avec le dos de la

langue abaissé. Concernant la palatalisée /tʲ/, pour la totalité des sujets, la palatalisation s'effectue grâce au geste d'élévation du dos de la langue au niveau de la bobine mid-dorsum. Pour la plupart des locuteurs, cette action d'élévation dorsale entraine probablement l'élévation de la lame de la langue. Concernant la position sur l'axe des abscisses, pour certains de nos sujets, /tʲ/ est plus antérieure que /t/ au niveau du post-dorsum. La figure ci-dessous représente les localisations des 4 bobines dans la cavité buccale ainsi que la trace du palais des 3 locuteurs types pour la paire minimale /tuk/~/tʲuk/ dans le plan X/Y du sujet. Ainsi, pour IM et MT /t/ est apicale avec une forme linguale plutôt aplatie pour IM et concave pour MT. La production de la palatalisée /tʲ/ montre une élévation des points lame et mid-dorsum pour les deux locutrices avec une langue bombée pour IM et plate pour MT. Pour MK, /t/ est apico-laminale tandis que la forme linguale lors de la production de /tʲ/ à l'atteinte de la cible est semblable à celle réalisée par IM. De plus, pour produire /tʲ/ on observe une élévation au niveau du post-dorsum pour les 3 sujets et une antériorisation de ce point pour IM et MT.

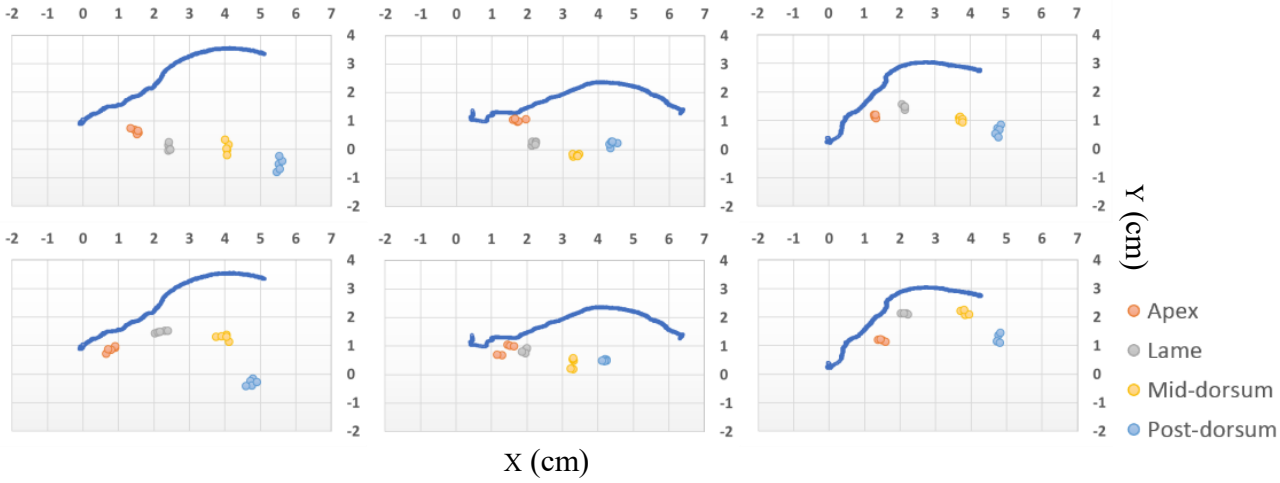



FIGURE 2 : Localisation des 4 bobines dans la cavité buccale des locuteurs IM (à gauche), MT (au milieu) et MK (à droite) dans la paire minimale /tuk/ (en haut) ~/tʲuk/ (en bas). La ligne continue indique le contour du palais.

Acoustiquement, la palatalisée /tʲ/ n'a jamais été réalisée [tʲ] par les locuteurs de notre étude qui l'ont majoritairement réalisée comme [ tʲ ]. Nous avons également relevé quelques spirantisations de /tʲ/ dans /tʲuk/ réalisées sifflante [s] ou chuintante [ʃ].

Des analyses statistiques ont été effectuées pour tester l'effet de l'accent et de la position dans la syllabe sur les coordonnées (X,Y) de chaque point de mesure. Les résultats obtenus pour la consonne /tʲ/ dans la paire /ˈtʲe.lo/ *vs* /tʲeˈla/ et les monosyllabes accentués /tʲap/ *vs* /matʲ/ montrent que l'accent et les positions de coda et d'attaque syllabique n'ont pas d'effet sur la forme linguale à l'instant de l'atteinte de la cible consonantique. Les mêmes résultats ont été obtenus pour les autres consonnes de notre étude /ʧ ʃ ʃː/. Pour celles-ci, de la même manière, aucune différence au niveau de la forme de la langue qui serait liée à l'accent ou à la structure syllabique n'a été relevée.

**Cas des consonnes /ʃ/ *vs* /ʃː/**

Nous avons relevé 4 stratégies articulatoires de réalisation des consonnes /ʃ/ et /ʃː/ en fonction des locuteurs. La figure 3 représente les localisations des 4 bobines dans la cavité buccale des 4 locuteurs types issus de l'ensemble des sujets enregistrés pour les consonnes /ʃ/ et /ʃː/ dans les mots /ʃov/ *vs* /ʃːelʲ/ dans le plan X/Y du sujet. Pour les locutrices MT et KB, /ʃ/ est apicale, en revanche pour MT la forme linguale est concave au niveau du dos de la langue (points mid- et post-dorsum) alors que pour KB la langue se creuse au niveau de la bobine lame. La locutrice IM produit un /ʃ/ apical avec une forme linguale bombée et l'arrière de la langue abaissé. Enfin pour la locutrice IH, la langue montre une configuration plate au niveau de la pointe, de la lame et du mid-dorsum avec le

post-dorsum légèrement plus bas. Concernant /ʃː/, nous avons observé une antériorisation du geste lingual au niveau des 4 localisations pour tous les locuteurs sauf pour IH pour laquelle on constate seulement un avancement du post-dorsum et une élévation au niveau des bobines apex et lame, le mid-dorsum fonctionnant comme un pivot. Pour la locutrice KB, on observe un abaissement de la partie antérieure de la langue à l'instant de l'atteinte de la cible /ʃː/. Pour les locutrices MT et IM, la palatalisation s'effectue en élevant lame, mid- et post-dorsum pour MT et lame pour IM avec une forme linguale aplatie pour la première et une langue bombée pour la deuxième.

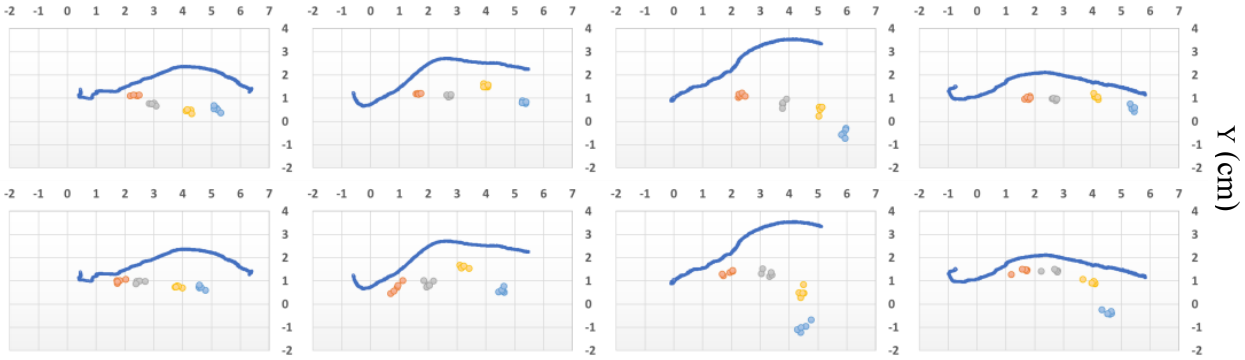



FIGURE 3 *:* Localisation des 4 bobines dans la cavité buccale des locuteurs MT, KB, IM et IH (de gauche à droite) dans /ʃov/ (en haut) *vs* /ʃː elʲ/ (en bas). La ligne continue indique le contour du palais.

Des configurations linguales similaires ont été trouvées pour /ʃ/ et /ʃː/ dans les mots /voʃ/ *vs* /lʲeʃː/, /ʃestʲ/ *vs* /ʃː it/ et /ʃes.ˈtʲi/ *vs* /ʃː i.ˈta/.

## 3.2 Patterns temporels

L'objectif est de décrire le timing du geste lingual pour atteindre les cibles /t tʲ ʧ ʃ ʃː/. Le timing est observé dans les trois modalités du facteur *renforcement articulatoire* à partir des points apex, lame, mid- et post-dorsum et calculé à l'aide des écarts temporels (Δt) entre le point minimal de la mandibule et le point maximal de chaque point localisé sur la langue. Cet intervalle temporel est considéré comme la mise en place du geste consonantique (cf. section 2.4).

**Modalité Accent*Onset**

Aucun écart temporel significatif n'a été trouvé entre les différentes parties de la langue lors de la production de /t/ et cela chez tous les locuteurs. Ce résultat suggère une activation simultanée des 4 zones linguales et ainsi un déplacement global de la langue. En revanche, pour /tʲ/ l'apex atteint sa cible dentale et/ou alvéolaire en premier suivie de l'élévation du mid- et du post-dorsum. Cet écart temporel entre l'articulation primaire apicale et l'articulation secondaire palatale pour /tʲ/ est trouvé significatif (z=4,49 p=.00148 pour mid- et z=5,27 p<.0001 pour post-dorsum). Les patterns spatiaux concernant la forme linguale lors de l'atteinte de la cible /tʲ/ suggèrent un rétrécissement au niveau pré- ou médio-palatal alors que pour /t/ le rétrécissement est présent au niveau de la région dento-alvéolaire. Nous avons comparé ensuite les variations de la variable *durée* en fonction des consonnes /tʲ ʧ ʃ ʃː/ toujours en position d'attaque dans la syllabe accentuée (fig. 4) en excluant volontairement /t/ pour les raisons évoquées *supra*. Pour /ʧ/, on retrouve à peu de choses près le même timing que pour /tʲ/ : l'élévation dorsale suit l'élévation apico-laminale (z=5,91 p=.0001 pour mid- et z=7,66 p<.0001 pour post-dorsum). Les patterns spatiaux de cette consonne montrent aussi qu'à l'instant de l'atteinte de la cible consonantique /ʧ/, les contours linguaux observés chez certains sujets suggèrent une articulation apicale ou apico-laminale. La forme linguale bombée au niveau du dos de la langue chez certains sujets peut être considérée d'après Ladefoged et Maddiesson (1996) comme *« a slight degree of palatalization »*. Pour /ʃ/ l'activation laminale est suivie de l'activation du post-dorsum (z=3,95 p=0,0171) et dans le cas de la fricative /ʃː/ aucun écart significatif entre les

4 points de référence n'a été observé. Concernant les différences interconsonnes, le déplacement des articulateurs apex et lame est significativement plus lent pour /ʃː/ que pour /tʲ ʧ ʃ/ : z=-6,47 p<.0001 ; z=-4,76 p=.00063 ; z=-5,22 p<.0001 pour l'écart concernant la localisation lame respectivement entre /ʃː/ et /tʲ ʧ ʃ/ et z=-6,38 p<.0001 ; z=-5,32 p<.0001 pour l'écart concernant la localisation apex respectivement entre /ʃː/ et /tʲ ʧ/. En effet, la production d'une fricative exige un contrôle plus fin de la constriction au niveau du lieu articulatoire pour générer le bruit de friction (Vallée et *al*., 2002). Comme /ʃː/ est une fricative phonologiquement longue, le contrôle doit être maintenu sur toute la durée consonantique. De plus, elle est palatalisée, et pour certains des locuteurs possède un double lieu articulatoire (ex. KB) ce qui exigerait un contrôle plus fin.

**Modalité Atone*Onset**

Les différences significatives entre les Δt moyens mesurés pour chacune des localisations sont moins nombreuses lorsque la consonne se trouve en attaque d'une syllabe inaccentuée (fig. 4). Pour /tʲ/ et /ʧ/ l'articulation apicale précède l'élévation du mid-dorsum pour /tʲ/ (z=4,16 p=.0075) et celle du post-dorsum pour /ʧ/ (z=4,27 p=.00402). Pour /ʃ/ et /ʃː/, l'activation des 4 points de mesure est simultanée. Au niveau interconsonantique un écart temporel significatif est trouvé pour la localisation lame entre /tʲ/ et /ʃː/ (z=-4,02 p=0,01261) et entre /ʃ/ et /ʃː/ (z=-3,70 p=0,04413).

**Modalité Accent*Coda**

Dans cette modalité nous n'avons pas relevé d'écart significatif dans le timing des 4 points de référence lors de la production d'une consonne donnée. Cela suggère une activation plus synchrone des 4 parties de la langue et donc un déplacement plus global de la masse linguale lorsque la consonne se trouve en position finale de syllabe. En revanche, quelques différences significatives qui n'ont pas été observées par Biteeva Lecocq et *al*. (2016) sont à signaler au niveau des comparaisons interconsonantiques (fig. 4). Ainsi, le geste apical est significativement plus lent pour /ʃ/ et /ʃː/ que pour /ʧ/ (z=-4,16 p=.00686 ; z=-5,19 p<.0001) et pour /ʃː/ que pour /tʲ/ (z=-4,34 p=.00319). Quant au mouvement dorsal, celui-ci est significativement plus lent, dans cette modalité, pour /ʃ/ et /ʃː/ que pour /ʧ/ au niveau du mid-dorsum (z=-4,38 p=.00301 ; z=-4,69 p=.00073) et que pour /tʲ/ au niveau du post-dorsum (z=-3,80 p=0,02871 ; z=-3,77 p=0,03295).

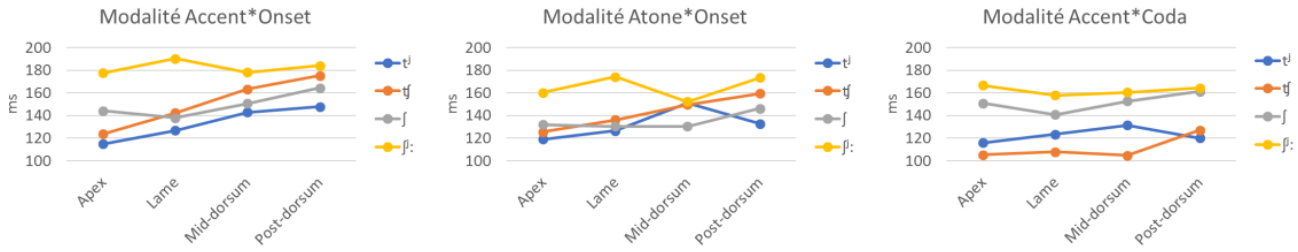



FIGURE 4 : Ecarts temporels moyens pour les 4 points de référence pour les consonnes /tʲ ʧ ʃ ʃː/ en fonction des 3 modalités du facteur *renforcement articulatoire*

# 4 Discussion

Les résultats obtenus montrent une variabilité des productions consonantiques qui relèvent davantage des locuteurs que de la structure syllabique ou de l'accent et rejoignent ceux obtenus par Biteeva Lecocq et *al*. (2016). Pour la totalité des locuteurs, la palatalisation s'effectue grâce au geste d'élévation du dos de la langue au niveau du mid-dorsum, action qui, probablement, provoque l'élévation laminale. Concernant l'axe antéro-postérieure, chez certains sujets, les palatalisées sont plus antérieures que les non palatalisées au point post-dorsum. Les différences sont trouvées au niveau de la forme globale de la langue qui peut être bombée chez certains sujets ou avoir une configuration plate avec une constriction formée par l'avant de la langue chez d'autres. Nos résultats rejoignent ceux de Kedrova et *al*. (2008) qui ont observé deux stratégies de réalisation des palatalisées du russe. La première engage massivement le corps de la langue et, plus précisément, présente une élévation importante du dos de la langue ainsi que son avancement dans la cavité buccale. La seconde implique davantage l'activité de la lame dans la région post-alvéolaire. La

proposition faite par Biteeva Lecocq et *al*. (2016) concernant une incidence de lieu d'articulation primaire sur l'articulation secondaire palatale n'a pas trouvé de confirmation dans cette étude. En effet, il a été suggéré que si l'apex se dirigeait vers les incisives inférieures dans /t/, il suffisait au dos de la langue de s'élever vers le palais pour produire /tʲ/. En revanche, si /t/ était produit avec la pointe relevée, le degré de palatalisation semblait moins important. Ces résultats suggéraient que que la pointe relevée dans la production d'une articulation initiale freinait le degré de la palatalisation. Les résultats obtenus à partir des 9 sujets observés ne montrent pas d'impact clair de lieu d'articulation initial sur l'articulation secondaire de type palatal. Nous nous attendions également à ce que les locuteurs originaires de régions géographiquement proches présentent des stratégies plutôt semblables de production. Même si quelques ressemblances au niveau des formes linguales pour une consonne donnée (chez les locuteurs des régions voisines) ont été trouvées, nous n'avons pas observé d'influence claire du contexte diatopique sur les productions. D'après nos données, la forme du palais n'a visiblement pas d'incidence sur la configuration du geste lingual à l'atteinte de la cible consonantique. En effet, certains participants dont les tracés du palais sont similaires emploient des stratégies différentes pour réaliser une consonne donnée.

Nous avons également relevé quatre stratégies articulatoires utilisées pour /ʃ/ : deux d'entre elles impliquent une forme concave de la langue au niveau du dos (MT) ou bien au niveau de la lame (KB). Ces réalisations ont déjà été décrites par Ladefoged et Maddieson (1996) mais pour la production de la sifflante /s/ par deux locuteurs de l'anglais britannique. En effet, celle-ci a été produite soit avec « *a deep hollow in the center of the tongue* », soit « *a hollowing of the tongue just behind the constriction* » (Ladefoged et Maddieson, 1996, p. 147). Cette dernière réalisation a également été observée pour /ʃ/ du russe par Akishina et Baranovskaya (2011). Pour la locutrice IM la forme linguale rappelle celle de la palato-alvéolaire /ʃ/ de l'anglais référencée toujours dans Ladefoged et Maddieson (1996) et décrite comme « *a post-alveolar domed sibilant*. »

Nos résultats concernant le pattern spatial à l'instant de l'atteinte de la cible ne montrent pas d'effet de l'accent ni d'effet de la position dans la syllabe pour l'ensemble des locuteurs enregistrés. Contrairement aux propositions de Straka (1963), Fougeron (1998), l'hypothèse concernant un geste lingual plus ample et plus stable sous l'influence du facteur *renforcement articulatoire* n'a pas été validée par nos observations. Nous avons observé l'effet de l'accent et de la position sur les patterns temporels du geste linguo-palatal plutôt que sur la forme de la langue à l'atteinte de la cible étudiée. Pour les modalités Accent*Onset et Atone*Onset, un écart temporel significatif est trouvé chez tous les locuteurs entre l'articulation primaire et secondaire. Ces observations rejoignent celles de Ladefoged et Maddieson (1996). Quant aux palatalisées et palatales /tʲ ʧ ʃ ʃʲ:/, dans la modalité Accent*Onset, des écarts temporels significatifs ont été trouvés dans la coordination temporelle des 4 points de mesure lors de la production d'une consonne donnée ainsi que dans le timing d'un même point de référence entre les différents types consonantiques. Ces écarts sont moins souvent significatifs dans la modalité Atone*Onset et sont observés seulement au niveau des comparaisons interconsonantiques dans la modalité Accent*Coda. La position coda, connue pour être une position de relâchement articulatoire, a tendance à niveler les différences dans la dynamique du geste lingual pour l'ensemble des consonnes étudiées. Ce résultat gagnera en précision avec un prolongement de l'analyse en intégrant la modalité Atone*Coda avec la structure /ˈCV.VC/ dans le facteur *renforcement articulatoire*. Ainsi, contrairement aux observations faites par Browman et Goldstein (1995), nos patterns spatiaux ne présentent pas une réduction de l'amplitude du mouvement lingual en coda qui serait causée par une baisse de l'effort articulatoire ou une coopération des différentes régions de la langue ayant pour but d'effectuer un geste moins extrême dans cette position. D'autre part, d'après Knyazev et Pozharitskaya (2012), l'accent en russe affecte plutôt le noyau vocalique. Les voyelles en syllabe accentuée sont généralement plus longues que lorsqu'elles se trouvent en dehors de l'accent tonique, position dans laquelle une diminution de l'amplitude du mouvement lingual et donc le phénomène d'*undershoot* sont observés.

# Remerciement

Soutien financier de l'IRS IDEX ComUE UGA PALGEST, du Conseil Européen de la Recherche sous le septième programme-cadre de l'Union Européenne (FP7/2007-2013 Grant Agreement no. 339152, "Speech Unit(e)s", PI: Jean-Luc-Schwartz) et ANR-10-BLAN-1916 APPSy.

# Références

AKISHINA A. & S BARANOVSKAYA (2011). *Russkaia fonetika na fone obshei*. Moscow: Librokom.

BAZZOLI C., LETUÉ F. & M.-J. MARTINEZ (2015). Modelling finger force produced from different tasks using linear mixed models with lme R function. *Journal of Case Studies in Business, Industry and Government Statistics (CSBIGS)* 6(1), 16-36.

BITEEVA LECOCQ E., VALLEE N., GERBER S. & C. SAVARIAUX (2016). Variabilité du geste palatal : effet du locuteur, de la structure syllabique et de l'accent sur différents types de consonnes en russe. *Actes de la conférence conjointe JEP-TALN-RECITAL 2016*, volume 1 : JEP.

BROWMAN C. P. & L. GOLDSTEIN (1988). Some notes on syllable structure in articulatory phonology. *Phonetica* 45, 140-155.

BROWMAN C. P. & L. GOLDSTEIN (1992). Articulatory phonology: An overview. *Phonetica* 49, 155-180.

BROWMAN C. P. & L. GOLDSTEIN (1995). Gestural syllable position effects in American English. In Bell-Berti F. & L.J. Raphael. *Producing Speech: Contemporary Issues*, 19-33. New York: AIP Press.

BYRD D. (1995). C-centers revisited. *Phonetica* 52, 285-306.

FOUGERON C. & P. KEATING (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America* 101, 3728-3740.

FOUGERON C. (1998). *Variations articulatoires en début de constituants prosodiques de différents niveaux en français*. Thèse de doctorat, Université Paris III, Paris. http://lpp.in2p3.fr/IMG/pdf/thesecfougeron-nonve_rifie_e.pdf. [consulté le 04/04/2015].

KAVITSKAYA D. (2002). Perceptual salience and palatalization in Russian. In Goldstein, L., Whalen D.H. & C. T. Best (Eds.), *Laboratory Phonology 8*, 589-610. Berlin: Mouton de Gruyter.

KEATING P. (1988). Palatals as complex segments: X-ray evidence. *UCLA Working Papers in Phonetics* 69, 77-91.

KEATING P. (1993). Phonetic representation of palatalization versus fronting. *UCLA Working papers in phonetics* 85, 6-21.

KEDROVA G. Y., ANISIMOV N. V., ZAHAROV L. M. & Y. A. PIROGOV (2008). Magnetic Resonance investigation of palatalized stop consonants and spirants in Russian. *Journal of the Acoustical Society of America* 123(5), 3325.

KELSO J. A. S., SALTZMAN E. L. & B. TULLER (1986). The dynamical perspective on speech production: data and theory. *Journal of Phonetics*, 14(1), 29-59.

KINGSTON J. (2008). Lenition. In *Selected proceedings of the 3rd conference on laboratory approaches to Spanish phonology* 1-31. Somerville, MA : Cascadilla Press.

KOCHETOV A. (2002). *Production, perception and emergent phonotactic patterns: A case of contrastive palatalization*. New York: Routledge.

KUZNETSOVA A. (1969). Nekotorye voprosy foneticheskoi kharakteristiki iavlenia tverdosti - miagkosti soglasnykh v russkikh govorakh. In S. Vysotskii (Ed.), *Eksperimentalno-foneticheskoe izuchenie russkikh govorov*, 35-215. Moscow: Nauka.

LADEFOGED P. & I. MADDIESON (1996). *The sounds of the world's languages.* Oxford: Blackwell.

LINDBLOM B. & I. MADDIESON (1988). Phonetic universals in consonant systems. In L. Hyman (Ed.), *Phonological acquisition and change*. New York: Academic Press.

MADDIESON I. (1984). *Patterns of sounds*. New York: Cambridge University Press.

RECASENS D. (1990). The articulatory classification of palatal consonants. *Journal of Phonetics* 18, 267-280.

RECASENS D., FARNETANI E., FONTDEVILA J. & M.D. PALLARÈS (1993). An electropalatographic study of alveolar and palatal consonants in Catalan and Italian. *Language and Speech* 36(2-3), 213-234.

RECASENS D., FONTDEVILA J. & M.D. PALLARÈS (1995). A production and perceptual account of palatalization. Phonology and phonetic evidence. In B. Connell & A. Arvaniti (eds.), *Papers in laboratory phonology IV*. Cambridge: Cambridge university press, 265-281.

RECASENS D. & J. ROMERO (1997). An EMMA study of segmental complexity in alveolopalatals and palatalized alveolars. *Phonetica* 54, 43-58.

SAGEY E. (1986). *The representation of features and relations in non-linear phonology*. Doctoral dissertation, MIT, Cambridge, Massachusetts.

SKALOZUB L. (1963). *Palatogrammy i rentgenogrammy soglasnykh fonem russkogo literaturnogo iazyka*. Kiev: s.n.

STRAKA G. (1963). La division des sons du langage en voyelles et consonnes peut-elle être justifiée ? *Travaux de linguistique et de littérature* 1, 17-99.

STRAKA G. (1965). Naissance et disparition des consonnes palatales dans l'évolution du latin au français. *Travaux de linguistique et de littérature* 3, 117-167.

VALLÉE N., BOË L.-J. & M STEFANUTO (1999). Typologies phonologiques et tendances universelles. Approche substantialiste. *Linx*, 11, 31-54. http://linx.revues.org/863. [consulté le 26 février 2015].

VALLÉE N., BOË L.-J., SCHWARTZ J.-L., BADIN P. & C. ABRY (2002). The weight of substance in phonological structure tendencies of the world's languages. *ZAS Papers in Linguistics* 28, 145-168. Berlin.

# Vers un modèle du « toucher vocal » pour la communication ubiquïte

Ambre Davat[1, 2] Véronique Aubergé[2] Gang Feng[1]
(1) Univ. Grenoble Alpes, Grenoble INP, GIPSA-lab, F-38040 Grenoble, France
(2) CNRS, Grenoble INP, LIG, 38000 Grenoble, France
ambre.davat@gipsa-lab.grenoble-inp.fr, veronique.auberge@univ-grenoble-alpes.fr, gang.feng@gipsa-lab.grenoble-inp.fr

## RESUME

Un des enjeux de la robotique de téléprésence est d'offrir une immersion socio-relationnelle ubiquïte. Pour y parvenir, il est nécessaire de comprendre et modéliser les facteurs qui permettraient au téléopérateur de contrôler la transmission de ses productions vocales, afin de lui en donner la perception, la proprioception et l'inter-proprioception. Ce modèle de transfert de la distance vocale socialement incarnée devra tenir compte de l'ensemble des paramètres impliqués dans l'effet socio-relationnel des productions vocales, en particulier l'intensité. Il devra également intégrer des éléments de contexte pertinents à la performance intentionnelle du locuteur téléopérant. Nous présentons dans cet article une première expérience visant à mesurer, analyser et modéliser la manière dont l'humain perçoit la distance physique qui le sépare d'un interlocuteur, en fonction de variations socio-affectives dans les productions vocales de cet interlocuteur. Ces résultats seront la référence des modèles qui seront implantés sur notre robot de téléprésence : RobAir Social Touch.

## ABSTRACT

**Towards a model of "social touch" for ubiquitous communication**

One of the challenges of telepresence robotics is to provide ubiquitous social-interpersonal immersion. In order to achieve this, there is a need to understand and model the factors that would allow the users to control the transmission of their vocal productions, and to give them perception, proprioception and inter-proprioception of this control. This model for transferring socially embodied vocal distance should take into account all parameters involved in the social-interpersonal effect of vocal productions, especially intensity. It should also integrate the background information which is relevant for the speakers to express their intentions. In this paper, we present a first experiment for measuring, analyzing and modeling how human beings perceive the distance to an interlocutor, depending on socio-affective variations in the vocal productions of this interlocutor. These results will be the reference for the models which will be implanted in our telepresence robot: Robair Social Touch.

MOTS-CLES : toucher social, robotique de téléprésence, psychoacoustique, perception, socio-affects, expérimentation écologique.

KEYWORDS: social touch, telepresence robotics, psychoacoustics, perception, socio-affects, ecological experimentation.


10.21437/JEP.2018-35

# 1 Introduction

Après l'invention de la téléphonie, qui permet de transporter la voix de quelqu'un d'un endroit à un autre, puis celle de la visiophonie, qui transmet également l'image, la robotique de téléprésence constitue une nouvelle étape dans l'immersion ubiquïte. En effet, il ne s'agit plus simplement de transporter la voix et l'image d'une personne vers un point de l'espace fixé par ses interlocuteurs, mais de permettre à l'utilisateur de contrôler ce point et de le déplacer dans l'espace distant dans lequel il communique à travers le « corps » physique d'un artefact robotique.

Ces robots de téléprésence reposent sous un nouvel angle la question de la fidélité avec laquelle la parole est transmise grâce aux moyens de télécommunication, question qui ne se pose plus pour le téléphone ou la visiophonie car nous avons culturellement intégré leurs artefacts dans nos usages. Ainsi par exemple, lors d'une communication téléphonique classique, les participants ont l'habitude de régler le volume sonore de leur interlocuteur au niveau qu'ils jugent acceptable. Une personne qui utilise un téléphone n'a donc pas à se préoccuper de savoir si sa voix est forte ou faible dans l'environnement distant car ce contrôle est clairement de la responsabilité de l'interlocuteur. En revanche, le pilote d'un robot de téléprésence peut être considéré par ses interlocuteurs comme maître de ce contrôle et donc responsable des artefacts liés à la télécommunication, d'autant plus que ses interlocuteurs sont réticents à l'idée de modifier les réglages du robot, car il est perçu non plus comme un simple objet de télécommunication, mais bien comme une personne physiquement présente.

Les systèmes proposés à l'heure actuelle dans la littérature pour tenter de résoudre ce problème ne nous satisfont pas entièrement. Ainsi, le feedback audio proposé par (Paepcke et al., 2011) peut effectivement inciter le téléopérateur à parler plus doucement, mais ne garantit pas que sa voix soit adaptée à l'environnement distant. Le système de (Takahashi et al., 2015) adapte automatiquement le volume sonore du robot en fonction du bruit ambiant et de la distance de l'interlocuteur, mais le contrôle de l'utilisateur est très limité, puisqu'il n'a le choix qu'entre deux modes de régulation : un mode « confortable » et un mode « secret » avec lequel seule une personne proche du robot peut l'entendre. Enfin, les interfaces visuelles proposées par (Kimura et al. 2007) permettent à l'utilisateur de visualiser le volume sonore perçu par son interlocuteur, mais semblent peu adaptées à la téléconférence mobile.

La difficulté à développer ce type de systèmes vient du fait que les productions vocales dépendent de multiples éléments de contexte (Cooke, 2014). Ainsi, le signal qui serait adapté dans l'environnement du téléopérateur, ne l'est pas forcément dans celui où se trouve le robot. Par exemple, si le téléopérateur se trouve dans une pièce calme, sa voix transmise par le robot ne sera peut-être pas suffisamment forte pour être audible dans un environnement bruyant. Il est donc intéressant de pouvoir l'amplifier artificiellement, en ayant conscience toutefois que cette correction peut engendrer des artefacts sonores, des voix « schizophoniques » qui ne pourraient pas être produites naturellement (Schafer, 1977). Or, cette schizophonie a sans doute un impact sur l'effet socio-relationnel des signaux vocaux. En effet, elle est notamment utilisée par la publicité, le cinéma et la musique afin de suggérer une certaine distance sociale entre l'auditeur et le locuteur (Maasø, 2008). Par exemple, un enregistrement qui parvient à capturer les bruits de bouche et la respiration du locuteur suggère une forme d'intimité à l'auditeur, qui ne devrait pas être capable de percevoir ces sons sans qu'ils soient extrêmement proches physiquement (Collins, Dockwray, 2015).

Pour améliorer notre robot de téléprésence, Robair Social Touch, nous cherchons à développer une interface qui permette à l'utilisateur de contrôler non seulement l'intensité de sa voix dans l'environnement distant, mais surtout son impact social, ou « toucher vocal ». Pour ce faire, nous avons besoin d'un modèle qui lie ce toucher vocal à un jeu de paramètres physiques mesurables (éléments de contexte et caractéristiques de la voix). Notre hypothèse est qu'il est possible de définir ce toucher vocal en étudiant la manière dont la perception acoustique de l'espace peut être influencée socialement. En effet, (Gardner, 1969), (Brungart, Scott, 2001) ou encore (Philbeck, Mershon, 2002) ont montré que la manière dont un sujet perçoit à l'aveugle la distance qui le sépare d'une source de parole dépend de l'effort vocal utilisé au moment de la production du stimulus. Ainsi, un chuchotement est systématiquement perçu plus proche qu'il ne l'est réellement, tandis qu'un cri est perçu comme plus éloigné. La réponse des sujets est donc influencée par la distance de communication imaginée par le locuteur au moment de l'enregistrement du stimuli, et non simplement par la distance physique des sources sonores. Il y a donc bien un biais social dans la perception acoustique de l'espace, sur lequel nous allons nous appuyer pour établir notre modèle. Dans ce papier, nous présenterons le protocole d'expérimentation que nous avons conçu puis mis en œuvre, et dont la complexité est relative à la complexité du problème posé. Enfin nous présenterons nos premiers résultats.

# 2 Scénario expérimental en « caméra cachée »

Le cœur de notre protocole consiste à demander à un sujet d'estimer la position spatiale de son interlocuteur, pendant que celui-ci exprime différents socio-affects. Cependant, il est fondamental que le sujet ne soit pas conscient que nous mesurons sa capacité à repérer l'espace physique en fonction de variations socio-affectives des stimuli, car alors il procéderait à des méta-traitements cognitifs qui ne sont pas ceux que nous souhaitons mesurer directement. Pour assurer cette démarche d'observation en situation écologique, nous avons dû monter un scénario de type « caméra cachée » et baser notre expérience sur une tâche prétexte. De plus, comme ce sont les performances de régulation sociale humaine qui nous intéressent, nous avons décidé, sur ces deux contraintes, d'accepter les variabilités de production des stimuli inhérentes à l'écologie naturelle de production des humains interactants. Ainsi, dans cette expérience, la source sonore n'est pas un haut-parleur, mais un locuteur expert, dont les productions vocales seront analysées a posteriori pour contrôler leur régularité.

En pratique, les sujets sont recrutés pour passer une expérience sur la perception du goût et de l'odorat. Cette expérience est censée faire partie d'un projet franco-japonais qui s'intéresse à la dimension culturelle et sociale des saveurs et des odeurs. Elle se déroule en binôme : un des sujets doit goûter des mini-pilules gustatives, l'autre respirer des boîtes à odeurs. Avant l'expérience, ils doivent remplir un questionnaire concernant leur pratique gustative et olfactive, et qui permet de recueillir des informations sur leur accent. Au moment où le sujet (S) se présente pour passer l'expérience, il rencontre le locuteur expert (L), présenté comme le second sujet de l'expérience. L se fait passer pour un nez professionnel, ce qui justifie qu'il puisse produire des énoncés à la fois très autoritaires et très hésitants. Un expérimentateur (E) présente l'expérience, en s'appuyant sur les questions posées par L afin de convaincre S qu'il passe bien une expérience sur le goût et l'odorat.

S comprend ainsi que chaque mini-pilule est assortie à une boîte à odeur. A chaque étape, L goûtera une mini-pilule disposée dans des gobelets répartis sur deux rangées de tables et annoncera ce qu'il a reconnu. Ensuite, E donnera une boîte à odeur à respirer à S, qui devra alors discuter avec L jusqu'à ce que chacun donne un avis définitif. Les deux participants seront placés dans différentes

situations d'interaction, soit disant pour observer des variations de leurs capacités perceptives. Afin d'éviter qu'ils puissent lire sur le visage de l'autre des informations de plaisir ou de déplaisir qui pourraient influencer leur perception, S devra porter un masque qui l'aveugle et dissimule le bas de son visage. Il passera donc l'expérience assis sur une chaise au centre de la pièce, à côté de E. Cependant, cette situation serait très inconfortable pour L, qui aurait l'impression de parler à quelqu'un qui l'ignore. Ainsi, pour s'assurer que S reste concentré sur sa tâche et rassurer L, on demande à S d'indiquer avant chaque boîte à odeur :

- la direction dans laquelle se trouve L (avant, gauche, derrière ou droite)
- sa distance (devant la première table, entre les deux tables ou derrière la deuxième table)
- son orientation (face au sujet, ou dos au sujet)

Par ailleurs, L reçoit un ordre de passage indiquant les positions des gobelets et leur numéro. Au dos de cette feuille se trouve les indications pour la vraie expérience, à savoir :

- la direction (avant, gauche, derrière ou droite)
- la distance (proche, milieu ou loin)
- l'orientation (face au sujet, ou dos au sujet)
- l'intensité à produire (faible ou forte)
- le socio-affect à communiquer (confiance autoritaire ou doute poli)
- le mot-clé à prononcer (ex : pomme, orange...)

Entre chaque étape, de la musique est diffusée par des hauts parleurs placés derrière les oreilles du sujet pour le faire patienter et dissimuler les pas du locuteur expert. Le sujet et le locuteur expert portent un micro serre-tête Sennheiser HSP 4 relié à un émetteur radio afin que leur échange soit enregistré.

FIGURE 1 : Photo du dispositif expérimental

A la fin de l'expérience, un débriefing est effectué. Il permet d'abord de vérifier que le sujet n'a pas deviné le but réel de l'expérience. Ensuite, la supercherie est dévoilée, et on s'assure que le sujet a bien compris les buts de l'expérience pour qu'il puisse donner son consentement éclairé.

Il s'agit d'une expérience lourde à monter, à mettre au point, puis à reproduire pour chaque sujet, puisqu'elle dure environ 1h30, introduction et débriefing compris. Il est donc important de noter que les 6 sujets déjà enregistrés, ainsi que plusieurs sujets préalables, non conservés tant que la mise au point n'était pas stabilisée, n'ont pas montré de signe d'ennui, de désintérêt ou de charge cognitive trop forte par rapport à la tâche prétexte ; en outre, nous n'avons observé a priori ni d'effet d'apprentissage, ni de dégradations des réponses des sujets au cours de l'expérience (nous le vérifierons statistiquement quand nous aurons plus de sujets). Nous allons à présent détailler la manière dont ce dispositif expérimental a été choisi.

# 3 Dispositif expérimental

L'expérience se déroule sur la plateforme d'expérimentation Domus, du LIG. Il s'agit d'une salle de 7.1 x 8.6 m, que nous avons aménagé avec des paravents pour former un espace carré de 7.1 x 7.1 m, soit 10 m en diagonale. Nous avons mesuré un temps de réverbération de 0.8 s[1]. C'est donc une salle particulièrement réverbérante, dans laquelle une simple mesure de l'intensité acoustique ne permet pas de deviner la distance d'une source sonore. Une photo du dispositif expérimental apparaît en Figure 1.

## 3.1 Choix des distances

Pour définir les distances mises en jeu dans cette expérience, nous nous sommes intéressés à l'incertitude avec laquelle un sujet estime la distance d'une source sonore. En effet, la psychoacoustique a montré que lorsqu'on demande à un sujet d'estimer à plusieurs reprises la distance d'une même source sonore, ses réponses varient. En choisissant des distances suffisamment proches les unes des autres, il est donc possible d'induire le sujet en erreur, et donc peut-être d'observer un biais socio-affectif dans ses réponses. Au contraire, si les distances mises en jeu sont trop différentes, le locuteur n'a aucune difficulté à deviner si son interlocuteur est devant, derrière, ou entre les deux tables. Or, cette incertitude dépend probablement des caractéristiques acoustiques de la salle où se déroule l'expérience, et varie d'un sujet à l'autre. Ainsi, (Zahorik et al., 2005) évoquent un flou perceptif évalué entre 5 et 20 % de la distance effective d'après un article de Haustein de 1969, et entre 20 et 60 % lorsque les distances sont représentées en échelle logarithmique d'après réanalyse de leurs propres travaux. Dans un autre article, (Calcagno, Abregú, 2012) indiquent l'écart type des distances estimées par leurs sujets. Celui-ci augmente linéairement dans le cas où les sujets doivent estimer la distance des sources sonores à l'aveugle. A titre indicatif, l'écart type est d'environ 40 cm pour une source située à 2 m, lorsque les sujets ont eu l'occasion de voir la salle avant le début du test. Par ailleurs, (Anderson, Zahorik, 2014) montrent que l'erreur de jugement des sujets suit une distribution normale lorsque la distance est représentée en échelle logarithmique.

Pour rester dans cette zone de flou perceptif, nous avons dû fabriquer de petites tables de 20 x 60 cm à partir de panneaux de bois posés sur des trépieds. Les 8 tables sont disposées sur 2 rangées, respectivement à 2 et 3 m de l'emplacement du sujet. Le locuteur se place donc à 1 m 70, 2 m 50 ou 3 m 30 du sujet. Non seulement cet aménagement convient parfaitement aux dimensions de la pièce, mais il est également intéressant en termes perceptifs. D'une part, il permet d'étudier le mode proche et le mode éloigné de la sphère sociale définie selon la théorie proxémique (Hall, 1966). D'autre part, les études psychoacoustiques ont montré que nous avons tendance à surestimer la distance des sources sonores proches et à sous-estimer celles des sources éloignées. La distance à laquelle s'inverse la tendance dépend des propriétés acoustiques de la pièce, mais semble être dans l'ordre de grandeur des distances que nous utilisons. Ainsi, (Anderson, Zahorik, 2014) ont observé un point d'inflexion à 1 m 90, puis à 3 m 22 dans une salle plus réverbérante, dans laquelle les distances sont perçues plus éloignées.

[1] Il s'agit du temps nécessaire pour que l'intensité du son diminue de 60 dB.

### 3.2 Choix des directions et orientations

Dans cette expérience, il ne s'agit pas de vérifier si nos sujets sont capables d'évaluer la direction d'arrivée des sons, mais de pouvoir éventuellement observer des biais perceptifs différents selon la manière dont le sujet et son interlocuteur sont positionnés dans l'espace. Parler en étant rigoureusement face à face avec quelqu'un, ce n'est pas la même chose que de lui parler de profil, ou même de dos. En effet, l'orientation du locuteur a d'abord une conséquence acoustique, car la voix humaine a une directivité : au lieu de se propager uniformément dans toutes les directions de l'espace, la puissance acoustique se répartit sous une forme cardioïde (Chu, 2002). En particulier, les hautes fréquences de la voix sont particulièrement atténuées derrière la tête du locuteur. En conséquence, un auditeur est capable de deviner à l'aveugle l'orientation de la tête d'un locuteur (Edlund et al., 2012). Par ailleurs, l'orientation du corps a un sens social ; elle est donc généralement prise en compte dans les études sur la proxémie, soit directement, par exemple comme dans (Remland et al.), soit indirectement, lorsque c'est le regard des interlocuteurs qui est analysé. Pour limiter la durée de l'expérience, l'orientation de L ne varie que lorsqu'il est entre les deux tables. Dans les autres cas, L est toujours tourné vers le sujet.

### 3.3 Choix des distances sociales des productions vocales

L doit être capable d'exprimer deux socio-affects différents : une confiance autoritaire par laquelle il marque une distance sociale grande et de dominance avec son interlocuteur, et un doute poli, par lequel il se rapproche socialement de son interlocuteur et inverse la dominance. Intrinsèquement à la nature prosodique de ces énoncés, L a tendance à parler plus fort pour exprimer la confiance autoritaire, et plus doucement pour exprimer le doute. Nous avons donc ajouté une consigne de contrôle de son intensité de production, faible ou forte, sur chacune des deux variables socio-affectives, afin d'observer si celle-ci a un impact sur la perception de l'auditeur.

## 4 Analyse des premiers résultats

Notre objectif est de faire passer 20 sujets sur cette expérience. Nous présentons ici les résultats obtenus avec nos 6 premiers sujets.

### 4.1 Vérification de la régularité de production du locuteur expert

Il est important de vérifier que L arrive à produire les différents socio-affects et intensités demandés, tant au niveau des contenus attitudinaux que des réalisations acoustiques.

Pour pouvoir faire une mesure étalonnée en décibels de l'intensité acoustique produite par L, nous aurions besoin d'un sonomètre placé à une distance fixe de sa bouche, ce qui est incompatible avec notre dispositif expérimental. Nous nous contentons donc de mesurer l'intensité des signaux numériques enregistrés durant l'expérience, ce qui est suffisant pour pouvoir les comparer entre eux. Pour ce faire, nous avons choisi la procédure suivante, réalisée par un script Praat :

1. extraction du pitch et de l'intensité du mot-clé (il s'agit bien sûre de l'intensité de l'enveloppe, et pas de l'intensité instantanée)
2. échantillonnage toutes les 10 ms du pitch et de l'intensité
3. moyennage des intensités pour lesquelles le pitch est défini

Les intensités mesurées à partir des enregistrements du micro porté par L sont présentées dans le Tableau 1. En moyenne, il y a bien un écart de plus de 7 dB entre les mots qui devaient être prononcés avec une intensité faible, et ceux devant être prononcés avec une intensité forte. Même si l'écart type ainsi que les valeurs min et max indiquent que l'intensité de certains mots-clés ne convient pas à leur catégorie, L arrive donc le plus souvent à suivre les consignes concernant les variations d'intensité. Cette méthode de mesure est néanmoins sensible à la position du micro, qui varie d'une expérience à l'autre, voire au cours d'une même expérience. En effet, ce micro n'est situé qu'à 2-3 cm de la bouche du locuteur, donc la moindre variation de son écartement fait varier l'intensité mesurée. Pour avoir accès à une seconde mesure d'intensité, nous avons ajouté un micro supplémentaire, placé à la verticale du sujet.

| Attitude | Doute poli | | Confiance autoritaire | |
|---|---|---|---|---|
| Intensité | Faible | Forte | Faible | Forte |
| Min | 49,6 | 57,8 | 51,1 | 58,9 |
| Max | 60,2 | 68,0 | 62,5 | 67,3 |
| Moyenne | 54,8 | 62,9 | 57,0 | 63,5 |
| Ecart type | 2,9 | 2,9 | 3,3 | 2,5 |

TABLEAU 1 : Mesure de l'intensité (dB) des mots-clés prononcés par le locuteur expert

Un test perceptif devra également être effectué pour confirmer que les attitudes sont bien reconnues, quelle que soit l'intensité utilisée. Nous n'avons pas encore fait d'analyse rigoureuse des différents mots-clés en terme de qualité de voix et de prosodie, mais nous avons déjà pu remarquer qu'il est difficile pour L d'exprimer une confiance autoritaire d'une voix faible et un doute poli d'une voix forte. A l'oreille, la stratégie qui semble efficace consiste dans un cas à parler très vite pour avoir une voix sèche, et dans l'autre à faire trainer les mots-clés et leur donner une tournure interrogative.

## 4.2 Premières observations

Une première manière d'analyser les réponses données par les sujets est d'étudier leur répartition en fonction des cinq variables étudiées à l'aide de matrices de confusion (Figure 2). Par manque d'espace, nous ne représenterons pas les résultats liés à la direction, car outre les confusions avant/arrière régulièrement observées en psychoacoustique, les sujets se trompent rarement pour estimer la direction. Pour simplifier la mise en page, les deux facteurs proxémiques et les quatre variables retenus sont présentés au même niveau, bien qu'a priori ils ne contribuent pas de façon équivalente au toucher vocal. Des tests du $\chi^2$ avec un seuil de tolérance de 5% ont été utilisés pour déterminer parmi ces huit matrices de confusion celles qui présentent des résultats significatifs, non explicables par le hasard.

Concernant la perception de la distance, on constate que les sujets perçoivent L comme plus éloigné lorsqu'il leur tourne le dos. En revanche, il n'y a pas d'effet notable du socio-affect ou de l'intensité. Par ailleurs, on note que les sujets sont indécis lorsqu'ils doivent estimer l'orientation de L, mais se trompent rarement lorsque L leur tourne le dos. Ils ont tendance à répondre que L leur tourne le dos lorsque le socio-affect indiqué à L est « doute poli », et l'intensité faible.

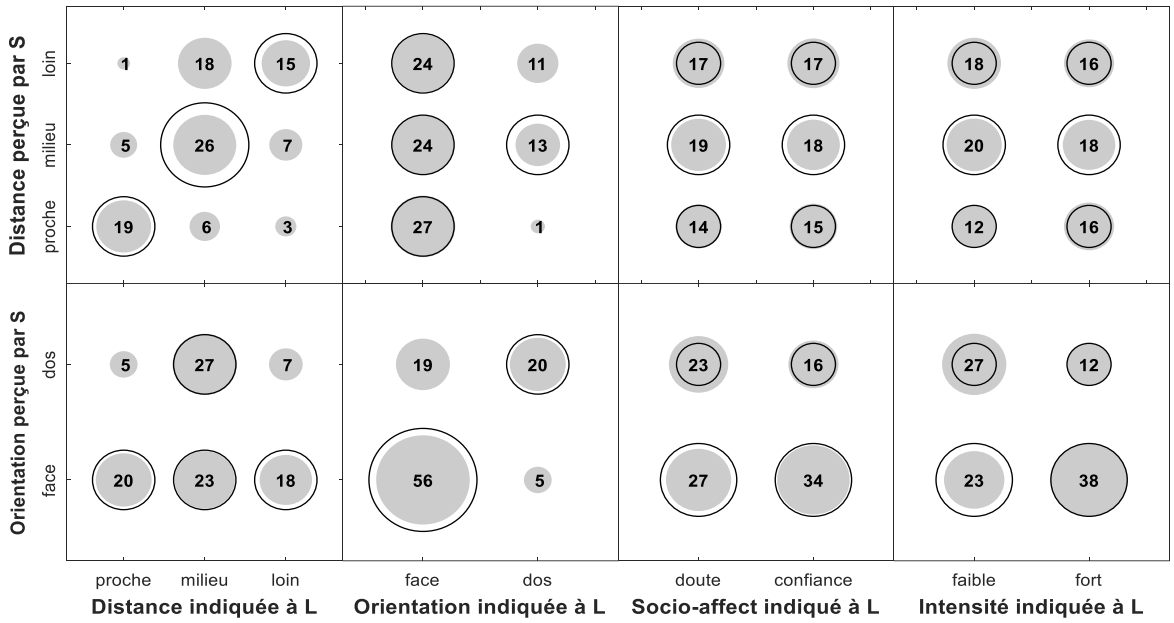



FIGURE 2 : Répartition des réponses données par les sujets (en %). *Le diamètre des disques gris est normalisé par rapport au pourcentage de réponses données pour une combinaison particulière d'indication donnée à L (indiquée en abscisse) et de réponse de S (indiquée en ordonnée). Les cercles noirs correspondent à la répartition qui aurait été obtenue si les sujets avaient donné la position réelle de L.*

# 5 Conclusion

L'interaction face à face in situ est un système complexe et fin qui permet aux interactants de se situer autant dans l'espace physique que dans l'espace social, la proxémie de leurs déplacements relatifs utilisant, différemment selon leurs cultures, cet espace physique pour signifier des informations sur leur espace social. Par cette étude, nous voulons montrer qu'il en est de même dans l'espace acoustique et que nous intégrons les distances physiques à nos productions vocales et notre audition pour exprimer et percevoir les distances socio-relationnelles. Ainsi, pour pouvoir assurer une immersion ubiquïte aussi bien physique que sociale, il faudra assister le téléopérateur dans la gestion de ces contrôles fins sous peine de générer des malentendus socio-relationnels importants. Nos premiers résultats montrent que le socio-affect exprimé par un locuteur et l'intensité de sa voix influencent la manière dont son orientation est perçue par un interlocuteur. Des analyses plus poussées devront être menées pour pouvoir établir un modèle du toucher vocal, qui sera testé et raffiné grâce à l'implémentation d'une interface pour robot de téléprésence.

# Références

ANDERSON P. W. & ZAHORIK P. (2014). Auditory/visual distance estimation: accuracy and variability. Frontiers in psychology, 5, 1097.

BRUNGART D. S. & SCOTT K. R. (2001). The effects of production and presentation level on the auditory distance perception of speech. The Journal of the Acoustical Society of America, 110(1), 425–440.

CALCAGNO E. R., ABREGÚ E. L., EGUÍA M. C. & VERGARA R. (2012). The role of vision in auditory distance perception. Perception, 41(2), 175–192.

CHU, W. T. ; SEARCH FOR : WARNOCK A. C. C. (2002). Detailed Directivity of Sound Fields Around Human Talkers. Rapport interne, Institute for Research in Construction (National Research Council of Canada, Ottawa ON, Canada). pp. 1–47.

COLLINS K. & DOCKWRAY R. (2015). Sonic proxemics and the art of persuasion: An analytical framework. Leonardo Music Journal, 25, 53–56.

COOKE M., KING S., GARNIER M. & AUBANEL V. (2014). The listening talker: A review of human and algorithmic context-induced modifications of speech. Computer Speech & Language, 28(2), 543–571.

EDLUND J., HELDNER M. & GUSTAFSON J. (2012). Who am i speaking at? Perceiving the head orientation of speakers from acoustic cues alone. In Proceedings of LREC Workshop on Multimodal Corpora for Machine Learning : LREC.

GARDNER M. B. (1969). Distance estimation of 0 or apparent 0-oriented speech signals in anechoic space. The Journal of the Acoustical Society of America, 45(1), 47–53.

HALL E. T. (1966). The hidden dimension.

KIMURA A., IHARA M., KOBAYASHI M., MANABE Y. & CHIHARA K. (2007). Visual feedback: its effect on teleconferencing. Human-Computer Interaction. HCI Applications and Services, p. 591–600.

MAASØ A. (2008). The proxemics of the mediated voice. Lowering the boom: critical studies in film sound, p. 36–50.

PAEPCKE A., SOTO B., TAKAYAMA L., KOENIG F. & GASSEND B. (2011). Yelling in the hall: using sidetone to address a problem with mobile remote presence systems. In Proceedings of the 24th annual ACM symposium on User interface software and technology, p. 107–116: ACM.

PHILBECK J. W. & MERSHON D. H. (2002). Knowledge about typical source output influences perceived auditory distance. The Journal of the Acoustical Society of America, 111(5), 1980–1983.

REMLAND M. S., JONES T. S. & BRINKMAN H. (1991). Proxemic and haptic behavior in three european countries. Journal of nonverbal behavior, 15(4), 215–232.

SCHAFER R. M. (1977). The tuning of the world. Alfred A. Knopf.

TAKAHASHI M., OGATA M., IMAI M., NAKAMURA K. & NAKADAI K. (2015). A case study of an automatic volume control interface for a telepresence system. In Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), p. 517–522: IEEE.

ZAHORIK P., BRUNGART D. S. & BRONKHORST A. W. (2005). Auditory distance perception in humans: A summary of past and present research. ACTA Acustica united with Acustica, 91(3), 409–420.

# L'opposition de voisement chez les apprenants syriens de FLE

Laura Abou Haidar
Laboratoire LIDILEM, univ. Grenoble Alpes, 38000 Grenoble, France
Laura.Abou-Haidar@univ-grenoble-alpes.fr

## RESUME

Dans cette étude nous analysons la réalisation de l'opposition de voisement des occlusives /p b t k/ chez des apprenantes arabophones syriennes de français langue étrangère, dans le cadre du projet IPFC-Arabe / *Interphonologie du français contemporain*. En nous basant sur 497 occurrences des phonèmes /p b t d/, produites dans le cadre des tâches de lecture et répétition du protocole IPFC par quatre locutrices syriennes, nous analysons la durée de l'occlusion, du VOT et de l'explosion, les configurations de voisement, et le V-ratio. Les résultats montrent une gestion spécifique des indices de voisement qui affecte non seulement la consonne /p/ dont on connaît les difficultés d'apprentissage mais aussi /b d/, le /t/ présentant le plus de stabilité. La consonne /p/ est affectée par un voisement prononcé alors que les consonnes /b/ et /d/ subissent un dévoisement notable qui est sans doute à interpréter comme une stratégie de surcorrection.

## ABSTRACT

**Voicing in Syrian learners of French as a second language**

We present an acoustical study of voicing of French /p b t d/ in female Syrian learners of French as a second language, within the IPFC-Arabic project. This study is based on the acoustical analysis of 497 occurrences of /p b t d/, produced by four female Syrian learners of French in two different tasks: reading and repetition of an isolated words list. We study some voicing distinctions as consonant closure, stop release, voicing configurations, and V-ratio. Results show that both /b d/ and /p/ are impacted by a specific management of voicing distinctions: /p/ is partially or totally voiced while /b d/ are partially or totally voiceless, which could be interpreted as a specific overcorrection learning strategy; /t/ shows most stability and is less impacted than other plosive oral consonants.

MOTS-CLÉS : Voisement – VOT – IPFC-Arabe – FLE – Apprenants syriens – L2.

KEYWORDS: Voicing – VOT – IPFC-Arabic – French as a second language – Syrian learners – L2.

# 1 Introduction

Si de nombreuses études explorent les corrélats acoustiques et perceptifs de l'opposition de voisement depuis des décennies (Lisker & Abramson, 1964 ; Abramson & Whalen, 2017), rares sont les travaux qui se sont intéressés à ce phénomène dans une situation d'apprentissage du français langue étrangère (désormais FLE), alors qu'il est couramment admis que l'acquisition de cette opposition constitue une difficulté majeure pour les apprenants de FLE. On peut citer le travail très complet de Landron (2017) qui effectue une recherche approfondie sur la production et la


10.21437/JEP.2018-36

perception du voisement des consonnes orales chez les locuteurs taïwanais, ou encore Birdsong (2003) dont l'échantillon est constitué d'apprenants anglophones tardifs de FLE.

La présente étude a pour objet l'opposition de voisement chez les apprenants arabophones de FLE, elle se situe dans le cadre du projet IPFC-Arabe (Abou Haidar et al., 2013) créé au sein d'IPFC-*Interphonologie du français contemporain* (Detey & al., 2010). Elle porte sur le versant production des consonnes orales voisées et non voisées de FLE chez des apprenantes arabophones syriennes. L'opposition de voisement des occlusives bilabiales orales, en particulier /p/ et /b/, est réputée être particulièrement problématique pour les locuteurs syriens de FLE, alors qu'aucune analyse n'est venue conforter d'une manière précise la nature des difficultés. Ce travail est une contribution à une meilleure connaissance des difficultés des apprenants arabophones syriens de FLE, en vue de la mise en place future d'un dispositif de remédiation adapté.

# 2 Protocole expérimental

## 2.1 Corpus et tâches

Cette étude porte sur 497 occurrences des phonèmes consonantiques du français /p b t d/ toutes locutrices confondues, correspondant à 205 réalisations pour /p/, 107 pour /t/, 116 pour /b/ et 69 pour /d/, en position initiale, médiane et finale de mot. Deux tâches du corpus IPFC-Arabe sont retenues : la répétition, sans support graphique, d'une liste de 64 mots, constituée pour moitié de mots appartenant au protocole commun PFC, et pour moitié de termes introduits en fonction des difficultés spécifiques des apprenants arabophones ; la lecture de la liste IPFC, conformément au protocole du projet.

| | |
|---|---|
| **/p/** | Parade – Peur – Pont – Pensons – Peu – Pan – Puce – Ponce – Port – Peureux – Pomper – Peau – Paresseux – Panse – Empoche – Epice – Tapisser – Apporte – Les pas – Pomper – Le pas - Lèpre |
| **/b/** | Bar – Balle – Boule – Bulle – Balade – Bombé – Bout – Visible – Zèbre – Bombé – Déductible - Tombe |
| **/t/** | Tante – Tant – Teint – Teinte – Tolérant – Content |
| **/d/** | Déductible – Dehors – Endurci – Andes – Parade - Inde |

TABLEAU 1 : Unités lexicales extraites de la liste spécifique IPFC-Arabe, sur lesquelles porte l'analyse

## 2.2 Locutrices

Quatre locutrices ont été retenues pour cette étude : des étudiantes syriennes, âge moyen de 25 ans (écart-type : 4), de niveau linguistique intermédiaire (B1-2 du CECRL), présentes en France depuis moins de 2 ans ; toutes suivaient la même formation linguistique de perfectionnement en FLE lors de la collecte de données (protocole global IPFC : lecture de deux listes de mots, répétition d'une liste, lecture d'un texte, conversation libre et guidée). Les données ont été collectées par des étudiantes de Master de FLE.

### 2.3 Méthode d'analyse

Le corpus a été segmenté et annoté sous Praat (Boersma & Weenink, 2018), dans le cadre d'une transcription semi-automatique avec le script Easyalign (Goldman, 2011), puis soumis à une analyse manuelle acoustique des indices suivants :

- La durée de l'occlusion en positions médiane et finale, dont la partie voisée ;
- Les « configurations du voisement » (Hallé & Adda-Decker, 2007) ;
- Le voice-ratio (Hallé & Adda-Decker, 2007), pourcentage rendant compte de la proportion de voisement sur la totalité d'un segment ;
- La durée du VOT (Lisker & Abramson, 1964), ou de l'explosion lorsqu'aucune installation postérieure du voisement n'était observée, en position finale par exemple.

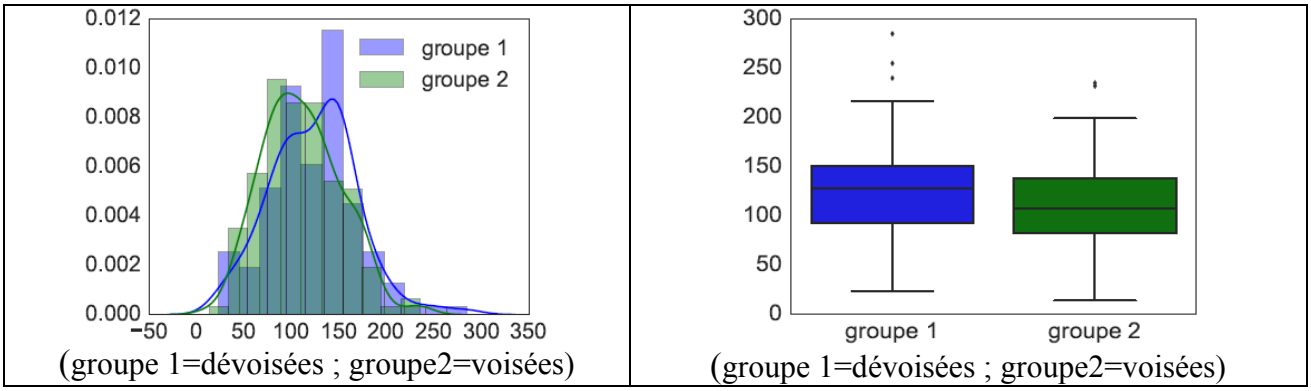
# 3 Résultats

### 3.1 Durée de l'occlusion

La distribution de la totalité des mesures de durée de la totalité des consonnes voisées et non voisées du corpus IPFC est représentée dans la figure 1. Avec une moyenne totale de 123ms (écart-type : 45) sur 145 occurrences de consonnes sourdes et 111ms (écart-type : 41) pour les 157 consonnes sonores, il apparaît que la durée de l'occlusion est un indice pertinent pour la distinction voisée/dévoisée chez ces locutrices (t-test, p=0,01). Les résultats des mesures de durée des réalisations phonétiques des consonnes occlusives /p b t d/ (toutes positions) sont représentés dans la table 2, en distinguant les tâches de lecture et de répétition.

| *Valeurs en ms* | Médiane | | | | Finale | | | |
|---|---|---|---|---|---|---|---|---|
| | /p/ | /b/ | /t/ | /d/ | /p/ | /b/ | /t/ | /d/ |
| **Moy-Lect** | **137** | **91** | **110** | **109** | **142** | **125** | **139** | **132** |
| ***Moy-Rép*** | ***133*** | ***102*** | ***96*** | ***107*** | ***60*** | ***80*** | ***119*** | ***129*** |
| σ-Lect | 51 | 22 | 29 | 94 | 98 | 26 | 32 | 40 |
| *σ-Rép* | *38* | *19* | *30* | *22* | *79* | *23* | *28* | *51* |

TABLEAU 2 : Mesures moyennes, minimales, maximales et écart-type des durées d'occlusion

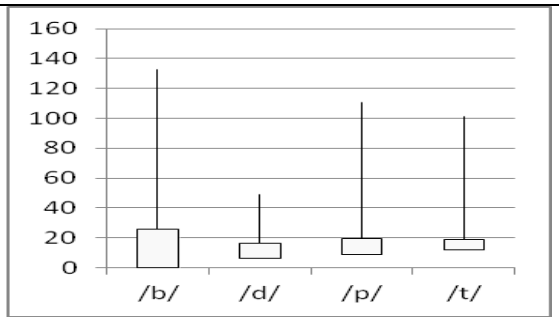



(groupe 1=dévoisées ; groupe2=voisées) (groupe 1=dévoisées ; groupe2=voisées)

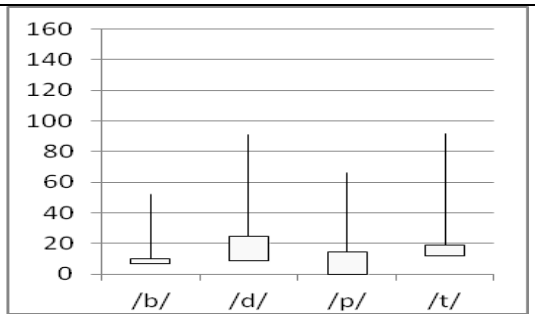
FIGURE 1 : Distribution des valeurs de durée (en ms) de *l'occlusion* des consonnes dévoisées (groupe 1) et voisées (groupe 2)

## 3.2 VOT et explosion

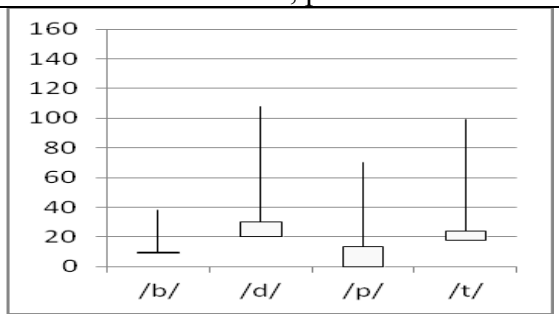
Les figures 2.1 à 2.6 représentent les mesures de VOT relevées en position initiale et médiane précédant l'installation de voisement (fig. 2.1, 2.2, 2.3, 2.4), et de l'explosion (fig. 2.5 et 2.6) relevée en position finale. Globalement on observe un effet de position important sur la valeur de l'explosion, avec une variabilité élevée des valeurs obtenues en position finale. Cependant, cette variabilité est à prendre avec une précaution extrême du fait des contextes segmentaux contraints de la liste IPFC qui n'était pas spécifiquement conçue pour une analyse fine de l'opposition de voisement. La consonne /t/ est celle qui semble présenter le plus de stabilité en position initiale et médiane. On observe également un effet de tâche sur les résultats, avec un impact aussi bien sur la dispersion des mesures que sur les durées maximales.



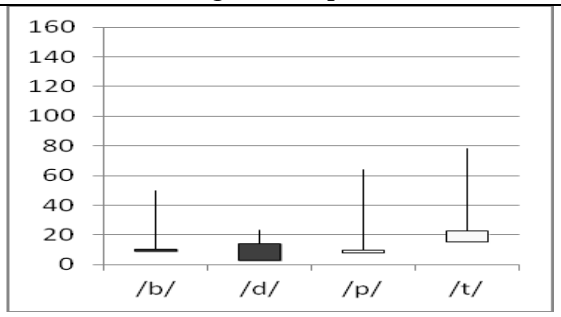

2.1 – Lecture, position initiale



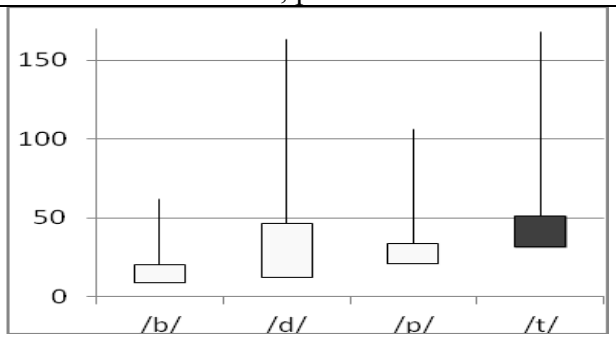

2.2 – Répétition, position initiale



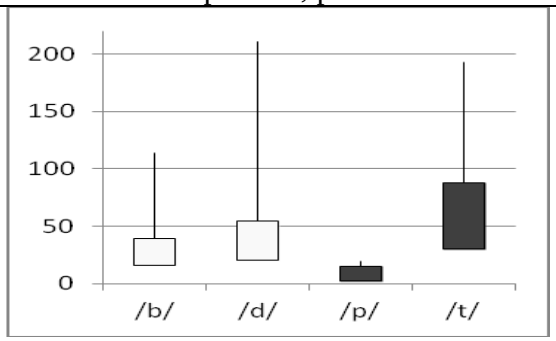

2.3 – Lecture, position médiane




2.4 – Répétition, position médiane




2.5 – Lecture, position finale




2.6 – Répétition, position finale

FIGURES 2.1-2.6 : Durées minimales, maximales, moyennes (en ms) et écart-type du VOT et de l'explosion

La durée de l'explosion est indice pertinent en production (Figure 3, t-test, p<0,0001) sur l'opposition de voisement, mais avec *un rapport inverse et non conforme à ce qui est habituellement admis pour le français* : l'explosion est plus longue pour les consonnes phonologiquement voisées que pour les consonnes phonologiquement non voisées. Une analyse plus fine à *contextes segmentaux strictement identiques* serait sans doute nécessaire dans une étape ultérieure, pour affiner, confirmer ou infirmer ces résultats : en effet, les consonnes voisées et non voisées du corpus sont présentes dans les deux groupes dans des contextes vocaliques et consonantiques variés (lieux d'articulation, aperture, nasalité, labialisation) dont il serait nécessaire de contrôler l'influence sur la durée de l'explosion.

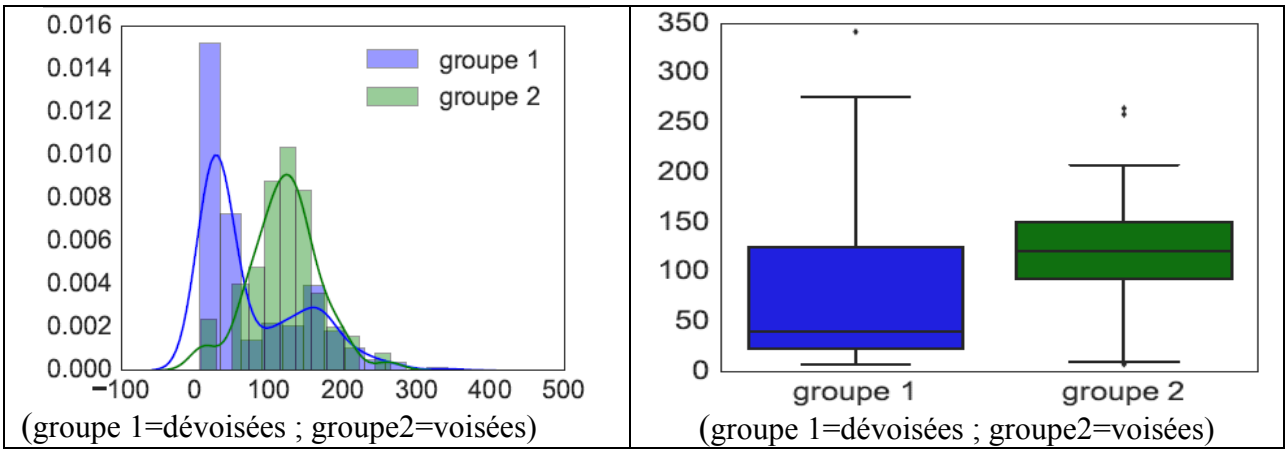
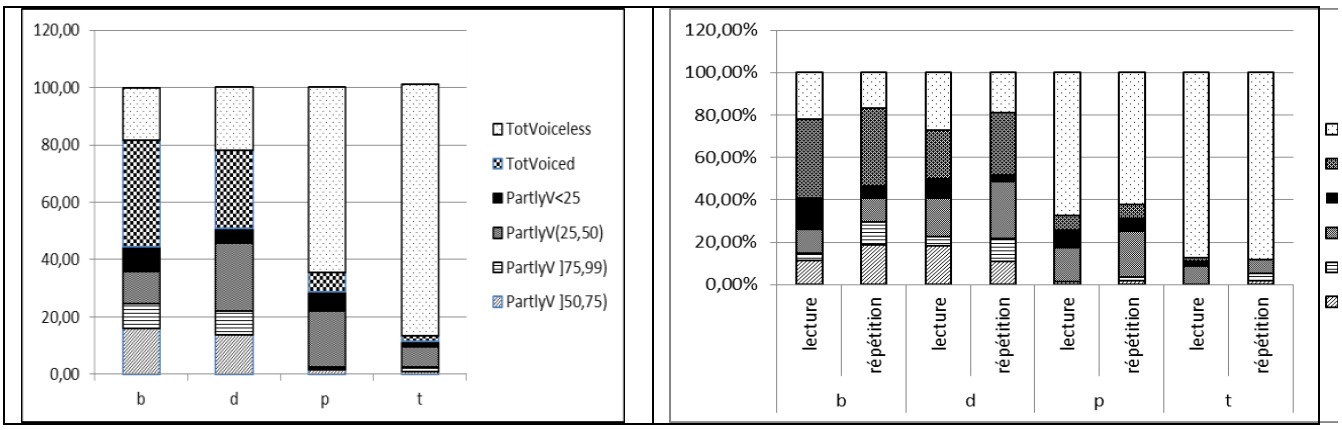



FIGURE 3 : Distribution des valeurs (en ms) de l'explosion des consonnes voisées et non voisées

## 3.3 Configurations de voisement

La figure 4 présente les différentes configurations de voisement (Hallé & Adda-Decker, 2007) pour toutes les réalisations phonétiques des consonnes phonologiques /p b t d/, toutes locutrices et toutes tâches confondues dans un premier temps. Nous avons distingué trois configurations parmi celles répertoriées par Hallé & Adda-Decker (2007, Fully voiced, Fully voiceless, et Partly voiced/bord gauche. A l'intérieur de cette configuration, nous distinguons 4 sous-catégories de voisement partiel (25% de voisement, 25% à 50%, 50% à 75%, et 75% à 99% de voisement). Les résultats confortent une des hypothèses principales relatives à la difficulté d'appréhender l'opposition de voisement des consonnes /p/-/b/ par les locuteurs syriens : 38% des occurrences correspondant à /p/ sont réalisées partiellement ou totalement dévoisées. Les mesures relevées pour les consonnes /b d/ sont plus inattendues : moins de la moitié des occurrences sont voisées, alors qu'on serait attendu à un voisement global sur ces consonnes qui appartiennent au système linguistique de la L1 des apprenantes. La consonne la moins affectée par ces altérations est /t/ dont plus de 83% des réalisations sont totalement dévoisées. L'effet de tâche est visualisé dans la figure 5. Les mesures relevées lors de la tâche de répétition permettent d'apprécier les capacités de production des apprenantes en fonction de leurs capacités de perception du caractère de voisement. En outre, les résultats de la tâche de lecture permettent d'apprécier les capacités de production, en fonction des capacités de décodage graphophonologique, indépendamment du processus de perception auditive. Les performances des apprenantes sont meilleures en répétition pour les consonnes voisées.




FIGURES 4 et 5 : Configurations de voisement toutes consonnes confondues dans le corpus (Fig. 4), et configurations de voisement en fonction de la tâche (Fig. 5)

## 3.4 V-ratio

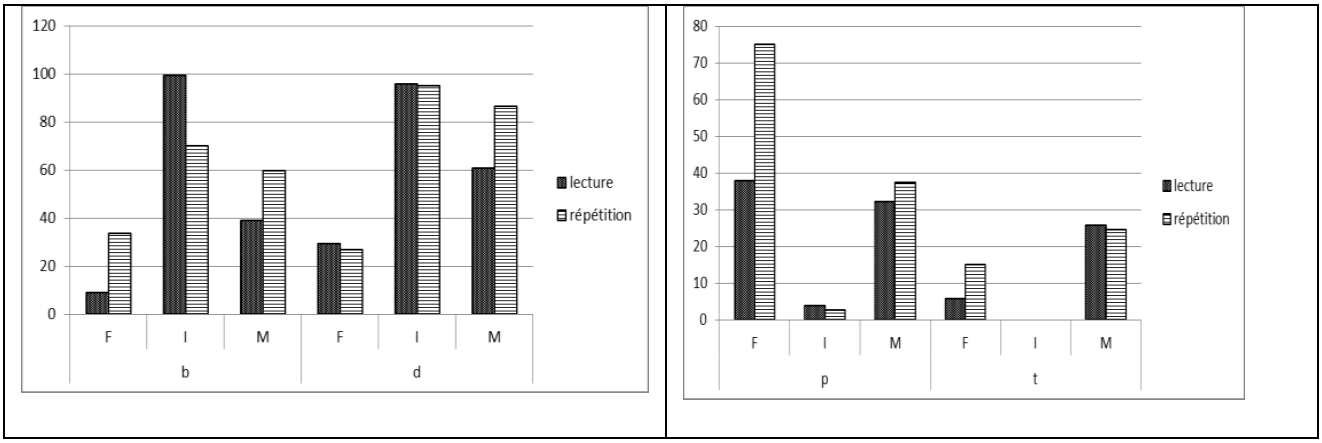
Les mesures de V-ratio (Hallé & Adda-Decker, 2007) relevées et représentées dans le tableau 3 montrent un voisement nettement moins massif qu'attendu sur les réalisations phonétiques de /b/ et /d/, avec des performances plus élevées pour la tâche de lecture (respectivement une moyenne de V-ratio de 62% et 56%) que pour la tâche de répétition (respectivement 55% et 47%). Pour /p/ on relève une moyenne de 18% avec une dispersion importante des valeurs. La consonne /t/ est celle qui présente le plus de stabilité.

| **Lecture** | /p/ | /t/ | /b/ | /d/ |
|---|---|---|---|---|
| *Min v-ratio/Lect* | *0* | *0* | *0* | *0* |
| *Max v-ratio/Lect* | *100* | *81* | *100* | *100* |
| *Moy v-ratio/Lect* | *18* | *6* | *62* | *56* |
| *Ecart-type v-ratio/Lect* | *29* | *17* | *38* | *38* |
| Min v-ratio/Rép | 0 | 0 | 0 | 0 |
| Max v-ratio/Rép | 100 | 100 | 100 | 100 |
| *Moy v-ratio/Rép* | *16* | *5* | *55* | *47* |
| Ecart-type v-ratio/Rép | 28 | 18 | 42 | 39 |

TABLEAU 3 : Mesures du v-ratio en fonction de la tâche

La prise en compte de la position du son dans l'unité montre des disparités très importantes des mesures de V-ratio (Figures 6 et 7). Le voisement de /b d/ se maintient prioritairement en initiale alors qu'on aurait pu s'attendre à une réduction liée à l'intensité qui accompagne cette position, avec des performances moins élevées en répétition qu'en lecture, ce qui peut révéler des difficultés spécifiques à l'identification perceptive de l'opposition de voisement. En position médiane le v-ratio présente des valeurs moins stables, qu'il serait nécessaire d'étudier d'une manière plus approfondie en fonction de la distribution et du voisinage vocalique. En position finale, /b/ et /d/ ont tendance à être fortement dévoisés. Pour /p/ et /t/, la position initiale conforte au contraire les effets attendus

d'une tension favorisant le dévoisement : une minorité de réalisations voisées pour /p/, et aucune pour /t/ dans cette position. Là encore c'est le /t/ qui présente le plus de stabilité avec des réalisations qui sont globalement moins soumises au dévoisement.




FIGURES 6 et 7: Moyenne du V-ratio pour les réalisations phonétiques de /b d/et /p t/ en fonction de la position et de la tâche

# Discussion

Des résultats inattendus ont été mis à jour dans cette étude concernant les consonnes occlusives orales /p b t d/. Les éléments qui ressortent confortent globalement la difficulté pour les apprenantes d'intégrer l'opposition phonologique de voisement à tel point que c'est le système dans son ensemble qui est déstabilisé, et pas seulement la réalisation de la consonne phonologique initialement absente de la L1, à savoir /p/. Globalement, la durée de l'occlusion n'est pas utilisée par ces 4 apprenantes syriennes comme indice pertinent pour différencier une voisée d'une non voisée correspondante. En revanche, la durée de l'explosion l'est, mais d'une manière inverse à celle qui est décrite dans la littérature pour le français : les consonnes phonologiquement voisées sont globalement affectées par un dévoisement partiel ou total, et l'allongement de la durée de l'explosion peut être interprétée comme un phénomène de surcorrection qui va amener l'apprenante à produire d'une manière altérée des phonèmes existant dans son répertoire langagier, et cela dans une tentative probable d'utilisation erronée, dans un contexte inapproprié, d'un trait phonologique en cours d'acquisition (à savoir l'opposition de voisement à travers entre autres un allongement de l'explosion des consonnes sourdes par rapport aux sonores correspondantes). Cela va dans le sens de l'élaboration d'un *système transitoire de l'apprenant* ou *interlangue* (Corder, 1980) avec un nouvel équilibre phonologique à trouver pour l'apprenant, et de nouvelles oppositions phonologiques à mettre en place, qui peuvent amener à une réduction du champ des réalisations antérieures et habituelles des phonèmes déjà présents dans le système linguistique de la L1 et du répertoire langagier existant. Comme prolongement de cette étude, il serait utile de confronter les résultats de l'analyse acoustique avec des tests perceptifs : en effet, la transcription « en aveugle » du corpus révèle des segments perceptuels inidentifiables ou non conformes à ce qui est attendu pour la réalisation de chacune des consonnes phonologiques du corpus. Une étude perceptive devrait permettre de préciser ou de pondérer le poids des indices relevés, en particulier pour ce qui est des configurations de voisement telles qu'elles ont été élevées dans cette étude. Il serait pertinent également d'enrichir le corpus en tenant compte de productions de locuteurs de sexe masculin, et de prévoir un groupe contrôle de locuteurs natifs. Il serait sans doute intéressant également d'intégrer

un indice acoustique complémentaire tel que le Voice Time Termination qui pourrait permettre d'affiner les résultats. Une adaptation du protocole expérimental par un contrôle strict du contexte segmental s'impose. Enfin, une approche comparative avec la gestion du voisement dans la langue maternelle, l'arabe syrien, devrait permettre d'affiner l'interprétation des résultats, afin de faciliter une remédiation approprié : nous pensons qu'enseigner la prononciation d'une LE consiste à intégrer, dans le répertoire langagier de l'apprenant, les spécificités de la langue cible ; ce qui veut dire comprendre la gestion de l'opposition de voisement dans la langue source, pour permettre à l'apprenant de mieux appréhender et mieux s'approprier les particularités de la langue cible.

# Remerciements

*Cette recherche a bénéficié d'une subvention allouée par la DGLFLF pour la constitution du corpus de locuteurs syriens dans IPFC-Arabe, au titre de l'action Maîtrise de la langue française.*

*Merci à Judith Abécassis pour son aide précieuse.*

# Références

ABOU HAIDAR L., ZEROUAL C., EMBARKI M., NABOULSI R. (2013). Projet IPFC-arabe : la variabilité des terrains de collecte. *Interphonologie du français contemporain : Corpus oraux en L2 et évaluation*. Journées IPFC-2013, Paris.

ABRAMSON A. S., WHALEN D. H. (2017). Voice onset time (VOT) : Theoretical and practical issues in measuring voicing distinctions. *Journal of Phonetics* 63, 75-86.

BIRDSONG (2003). Authenticité de prononciation en français L2 chez des apprenants tardifs anglophones : analyses segmentales et globales. *Acquisition et interaction en langue étrangère* 18, 2-14.

BOERSMA P., WEENINK D. (2018). Praat : doing phonetics by computer [Computer program]. Version 6.0.37, retrieved 3 February 2018 from http://www.praat.org/.

CORDER S.P. (1980). La sollicitation des données d'interlangue. *Langages* 57, 29-37.

DETEY S., DURAND J., LAKS B., LYCHE C. (2010). *Les variétés du français parlé dans l'espace francophone. Ressources pour l'enseignant.* Paris : Editions Ophrys.

GOLDMAN J.-P. (2011). EasyAlign: an automatic phonetic alignment tool under Praat. *Interspeech'11, 12th Annual Conference of the International Speech Communication Association*.

HALLÉ P., ADDA-DECKER M. (2007). Voicing assimilation in journalistic speech. At *Proceedings of the 16th International Congress of Phonetic Sciences*. Saarbrücken: Trouvain Jurgen / Barry William John (ed.), 493-496.

LANDRON (2017). *L'opposition de voisement des occlusives orales du français par des locuteurs taïwanais*. Thèse de Doctorat, université Sorbonne Nouvelle-Paris 3.

LISKER L., ABRAMSON A.S. (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word* 20, 384-422.

# Gestes et prosodie dans la parole aphasique non fluente

Gaëlle Ferré
LLING - UMR 6310, Chemin de la Censive du Tertre, BP 81227, 44312 Nantes cedex 3
Gaelle.Ferre@univ-nantes.fr

RÉSUMÉ

L'aphasie non-fluente se caractérise par de fréquentes recherches lexicales et un débit de parole beaucoup plus lent que la parole non aphasique. Pour les patients atteints de ce type d'aphasie, la communication avec leur entourage est de ce fait rendue difficile et se trouve souvent très altérée. Une des propositions thérapeutiques pour améliorer la qualité de vie de ces patients consiste à rééduquer les patients avec des alternatives plus multimodales. Cela suppose bien entendu que la gestualité soit une alternative possible pour les patients, et que leurs gestes ne soient pas affectés au même titre que leur parole. Cet article se propose donc d'étudier la gestualité de personnes aphasiques et de la comparer aux gestes produits par des personnes non aphasiques. Les résultats montrent que si l'on compare les gestes référentiels et non-référentiels, c'est essentiellement dans leur structure interne que les gestes des personnes aphasiques sont affectés, et que cette structure interne correspond aux troubles de la parole.

ABSTRACT

**Gesture and prosody in non-fluent aphasic speech.**

Non-fluent aphasia is characterized by frequent word searches and a slower speech-rate than is found in fluent speech. For people with aphasia (PWA), everyday communication is therefore very much impaired. Multimodal communication has been suggested as an alternative therapeutic treatment in order to improve patients' daily interactions. Such a treatment however presupposes that gesture is a possible alternative for patients, and that their gestures are not impacted by aphasia in the way their speech is. This paper therefore presents a study of PWA's gestures in a comparison with speakers without aphasia. Results show that PWA's referential and non-referential gestures are impacted essentially in their internal structure and that this structure matches PWA's speech impairment.

MOTS-CLÉS : Aphasie non fluente, prosodie, gestualité.

KEYWORDS: Non-fluent aphasia, prosody, gesture.

## 1 Introduction

L'aphasie est un trouble du langage entraîné par des lésions cérébrales, le plus souvent dues à un AVC, qui peut affecter la production linguistique écrite et orale et/ou la compréhension de la parole (Dipper *et al.*, 2015). Les lésions neuronales affectent des fonctions linguistiques spécifiques, les troubles pouvant révéler des déficiences phonologiques, syntaxiques, lexicales et sémantiques (Preisig *et al.*, 2015). La parole aphasique non-fluente se caractérise principalement par un effort de parole important lié à de fréquentes recherches lexicales, une articulation des sons difficile et une syntaxe


10.21437/JEP.2018-37

basée essentiellement sur la production de groupes nominaux (Macauley & Handley, 2005). Du fait des difficultés à communiquer qu'elle entraîne, l'aphasie a de lourdes conséquences psycho-sociales pour les personnes qui en sont atteintes (Blom Johansson, 2012; Nyström, 2006) :

— Perte d'emploi,
— Isolement des personnes aphasiques (PA) et dégradation des relations familiales et sociales,
— Perte d'estime de soi face à un sentiment d'incompétence linguistique pouvant mener à un état dépressif.

Afin d'améliorer la qualité de vie des patients, les thérapeutes et les chercheurs s'interrogent sur l'opportunité d'une thérapie multimodale (M-MAT : *Multimodal Aphasia Therapy*) en comparaison avec une thérapie centrée uniquement sur la parole (CIAT : *Constraint-Induced Aphasia Therapy*). Les résultats des études menées sur les progrès des patients ayant suivi l'une ou l'autre thérapie sont mitigés (Rose *et al.*, 2015) : les effets d'une thérapie multimodale dépendent de la situation de communication et de la sévérité de l'aphasie des patients. L'efficacité d'une thérapie multimodale présuppose d'ailleurs que la gestualité des patients ne soit pas affectée au même titre que la parole et puisse venir compenser la parole déficiente ou au moins faciliter l'accès au lexique. Si les travaux sur la gestualité produite – spontanément ou à la demande des thérapeutes – par les personnes aphasiques (PA) se sont multipliés ces dix dernières années, nous sommes encore loin de connaître les similarités et différences entre la gestualité des PA et celle des personnes non aphasiques (PNA), ni le rôle potentiellement compensatoire ou facilitateur des gestes produits par les PA (Pritchard *et al.*, 2015). Par ailleurs, les travaux existant se basent principalement sur des corpus de parole spécifiques qui peuvent avoir une grande influence sur la production gestuelle spontanée (narrations à partir d'images, production de discours procéduraux, …), ou sur des études qualitatives de la production de gestes par les PA dans des corpus non contrôlés (interactions en milieu familial).

L'objectif de cet article est donc de présenter une étude prosodique de la gestualité des PA dans des récits personnels semi-contrôlés en comparaison avec les gestes produits par des PNA : la distribution gestuelle des PA est-elle semblable à celle des PNA, le nombre de gestes et leur structure interne sont-ils similaires chez les deux groupes de locuteurs ?

# 2 Contexte théorique

Chez les aphasiques de Broca, les mots produits en position initiale de groupe sont plus longs que les mots produits en position finale à l'inverse du rythme de parole non aphasique (Danly & Shapiro, 1982) et les syllabes non-finales sont aussi plus longues que les syllabes finales (Louis, 2003). Les PA non fluentes montrent en particulier des difficultés à initier la production d'un mot (Kurowski & Blumstein, 2016), avec des erreurs phonologiques plus fréquentes en début de mot (Tuller, 1984). Les PA non fluentes souffrent également de nombreuses paraphasies : métathèses, épenthèses, substitutions et suppressions de son(s) dans les mots, assimilations, confusions de mots qui engendrent de nombreuses auto-corrections.

Les travaux menés sur la gestualité des PA trouvent parfois des résultats contradictoires. Selon Cicone *et al.* (1979) et Feyereisen (1983), les PA non fluentes ne sont pas plus fluentes en gestualité qu'elles ne le sont dans la parole. Elles produisent moins de gestes que les non-aphasiques, le débit

de parole étant directement corrélé à la fréquence gestuelle et au rapport geste/mot. D'autres travaux trouvent au contraire que la plupart des PA ont tendance à utiliser la communication non verbale pour compenser la parole déficiente (Hogrefe *et al.*, 2013; de Beer *et al.*, 2017). Cependant, parmi les travaux qui relatent une compensation possible de la parole par la gestualité, les résultats sont également contradictoires. Alors que pour Smith (1987) citée dans (Ahlsén, 1991), les PA les plus sévèrement atteintes compensent plus que les PA moins sévèrement atteintes, Mol *et al.* (2013) trouvent au contraire que les PA les plus sévèrement atteintes gestualisent moins facilement que les PNA pour compenser la parole. Ces différences dans les résultats obtenus pour la densité gestuelle des PA s'expliquent en grande partie cependant par le fait que les calculs n'ont pas été menés sur les mêmes unités de parole (Cocks *et al.*, 2013) : alors que certaines études choisissent le rapport geste/mot ou la proposition syntaxique pour calculer le nombre de gestes produits par les PA, d'autres choisissent une unité temporelle (nb de gestes par seconde ou minute).

Enfin, il n'est pas encore certain du rôle joué par la gestualité chez les PA : si certains auteurs comme ceux déjà cités dans le paragraphe précédent estiment que la gestualité peut être utilisée par les PA pour compenser une parole déficiente, d'autres pensent que la gestualité facilite la récupération de la forme phonologique correcte des mots dans les recherches lexicales sans se substituer à la parole (de Ruiter, 2006), quand d'autres encore accordent les deux fonctions aux gestes dans cette population (Kroenke *et al.*, 2013). Comme le remarque Ahlsén (2015), la gestualité linguistique et la parole sont étroitement liées, mais ont néanmoins aussi une certaine indépendance l'une par rapport à l'autre, ce qui expliquerait ces résultats en apparence contradictoires. Quant aux effets des thérapies multimodales, elles sont elles-aussi contrastées : si elles sont efficaces pour certains patients, elles ne le sont pas pour d'autres (Beeke *et al.*, 2015).

# 3 Questions de recherche

La plupart des études mentionnées dans la section précédente ont examiné les relations gestes/parole dans des expériences basées sur des descriptions d'images ou d'actions qui encouragent la production de gestes iconiques ou de pointages, les autres types de gestes ayant été largement ignorés. Nous connaissons donc mal la production gestuelle des PA dans des environnements moins contraints en comparaison avec les gestes des PNA. C'est sur ce point que cet article se propose de porter : lorsqu'ils ne sont pas contraints par la nature iconique d'une réponse ou par son lien avec un support imagé, la gestualité des PA est-elle similaire à celle des PNA ? De plus, la structure interne des gestes produits par les PA est-elle semblable à celle des PNA ? Par exemple, qu'advient-il des gestes lorsque les mots sont difficiles à exprimer ? Sont-ils maintenus, répétés ou abandonnés ?

# 4 Données et méthodologie

## 4.1 Corpus

Afin de répondre à ces questions, j'ai utilisé certains fichiers vidéos d'AphasiaBank (MacWhinney *et al.*, 2011), une base de données multimédias qui rassemble des locuteurs aphasiques et non-aphasiques de plusieurs langues, enregistrés selon le même protocole. Les tâches du protocole

incluent quatre genres : récits personnels élicités par des questions très similaires chez les PA et les PNA, descriptions d'images, récits de contes et discours procédural mais seuls les récits personnels ont été utilisés pour cette étude.

Dans la base de données, j'ai sélectionné 4 PA (2 femmes et 2 hommes) et 4 PNA (3 femmes et 1 homme), tous locuteurs d'anglais américain. Parmi les PA, 2 patients sont atteints d'une aphasie de Broca et les deux autres souffrent d'une aphasie transcorticale motrice. L'aphasie des quatre PA est modérée selon le score WAB (*Western Aphasia Battery*) qu'elles ont obtenu à une série de tests de compréhension et de production linguistique orale et écrite pratiqués en milieu hospitalier et répertoriés dans la base de données. Les informations concernant les quatre PA de l'étude sont données dans la Table 4.1.

| **Participants** | Sexe | Durée de l'aphasie | Date de l'enreg. | Âge | Aphasie | Score WAB |
|---|---|---|---|---|---|---|
| ACWT01a | F | 11 ans 8 mois | 2012 | 69,9 | Broca | 63,9 |
| ACWT02a | F | 3 ans 30 jours | 2012 | 53,1 | TCM | 74,6 |
| adler18a | H | 5 ans 75 jours | 2010 | 71,5 | TCM | 59,8 |
| elman03a | H | 11 ans | 2009 | 55,2 | Broca | 66,2 |

TABLE 1 – Description des PA sélectionnées dans AphasiaBank pour l'étude (les informations concernant la durée de l'aphasie et l'âge du participant sont données en relation avec la date où l'enregistrement a eu lieu)

## 4.2 Traitement des données

La nature du corpus pose la question des unités choisies pour comparer de manière adéquate la parole des PA et des PNA, sachant que les unités prosodiques ou syntaxiques traditionnellement adoptées pour la transcription dans les études multimodales peuvent se révéler problématiques. Par exemple, les Unités Inter-Pausales sont difficiles à adopter ici du fait de la présence récurrente de pauses silencieuses longues – y compris parfois en milieu de mot – chez les PA lors de leurs fréquentes recherches lexicales. Ces pauses fréquentes et parfois assez longues, ainsi que la présence de nombreuses amorces de mots rendent également difficile la détection de groupes intonatifs. Enfin, l'agrammaticalité de la parole des aphasiques de Broca (notamment l'absence récurrente de verbes et la substitution de certains groupes syntaxiques par des onomatopées, par exemple) exclut une transcription des propositions syntaxiques. Sachant que le phonème n'est pas une unité adaptée à une analyse multimodale car elle n'a pas du tout la même granularité que le geste, c'est le groupe de souffle qui a été choisi ici pour une transcription verbatim du corpus sous PRAAT (Boersma & Weenink, 2009). Chaque groupe de souffle était délimité par une reprise de souffle audible et/ou un changement de tour de parole. 224 groupes de souffle ont été annotés pour les PA contre 324 pour les PNA. Le corpus a ensuite été transcrit en mots également alignés avec le signal (694 mots pour les PA contre 2899 pour les PNA). Ces transcriptions incluaient les pauses silencieuses et remplies, ainsi que les amorces de mots, rires, etc. Dans une piste séparée, j'ai noté le nombre de syllabes dans chaque groupe de souffle, sans noter la transcription de chaque syllabe.

Avec le logiciel d'annotation des fichiers multimédias ELAN (Sloetjes & Wittenburg, 2008), l'ensemble des gestes manuels produits par les PA et les PNA sur les 8 récits personnels ont été

annotés (434 gestes pour les PA contre 291 pour les PNA, soit un total de 725 gestes). Les gestes ont été catégorisés en gestes référentiels (gestes qui font référence à un objet ou une action concrèt(e) ou à un concept abstrait) et en gestes non-référentiels (gestes qui ne font pas référence à un objet, une action ou un concept, comme par exemple, les gestes liés à la recherche lexicale, les gestes qui expriment une modalité ou qui sont liés à l'organisation du discours, ainsi que les battements). Au total, les PNA ont produit 132 gestes référentiels contre 159 gestes non-référentiels, et les PA 164 gestes référentiels contre 270 gestes non-référentiels.

Enfin, toujours sous ELAN, les différentes phases gestuelles (Kendon, 2004) ont été annotées. Parmi ces phases, on compte la *préparation*, c'est-à-dire la mise en place des mains pour la réalisation du geste. La phase de *réalisation* est la partie communicative du geste, et la *rétraction* correspond au relâchement des mains (ici, soit accompagné d'un retrait total vers la position de repos, e.g. les mains posées sur la table, soit accompagné d'un retrait partiel, e.g. relâchement des mains sans retrait total). Deux autres phases ont également été notées : une possible tenue *pré-réalisation* ou *post-réalisation* pendant lesquelles la/les main(s) marque(nt) une pause avant ou après la phase de réalisation, mais en conservant la configuration manuelle établie pendant la préparation ou pendant la réalisation.

# 5 Résultats

Afin de répondre aux questions posées dans la section 3, les différences entre les PA et les PNA ont été testées par Modèles Mixtes Linéaires Généralisés (GLMM) (Bates *et al.*, 2014) avec le programme d'analyses statistiques R 3.4.0 (R Core Team, 2012). Les locuteurs ont été inclus comme facteur aléatoire pour rendre compte de la variabilité inter-locuteurs.

Comme l'on pouvait s'y attendre, on observe une différence significative entre les participants (PA vs. PNA) sur la durée des groupes de souffle ($\beta$ = 1061.4, SE = 431.2, $p < 0.05$), avec une durée moyenne de 4029.7 ms pour les PA et de 2839 ms pour les PNA. On observe aussi une différence significative entre les deux groupes de participants pour ce qui est de la vitesse d'articulation (nb de syll/sec dans les groupes de souffle sans compter les temps de pause) avec en moyenne 2.5 syll/sec pour les PA contre 4.1 syll/sec pour les PNA ($\beta$ = -2.14, SE = 0.28, $p < 0.001$). Le débit de parole (nb de syll/sec en comptant les pauses) est lui-aussi significativement différent entre les deux groupes ($\beta$ = -2.3, SE = 0.24, $p < 0.001$) avec un débit moyen de 1.5 syll/sec pour les PA contre 3.8 syll/sec pour les PNA. Cette différence de débit n'est cependant pas liée à un nombre plus élevé de pauses remplies au sein des groupes de souffle pour les PA ($\beta$ = 0.63, SE = 0.29, $p$ = 0.27) ou de pauses silencieuses ($\beta$ = 1.40, SE = 0.33, $p$ = 0.14).

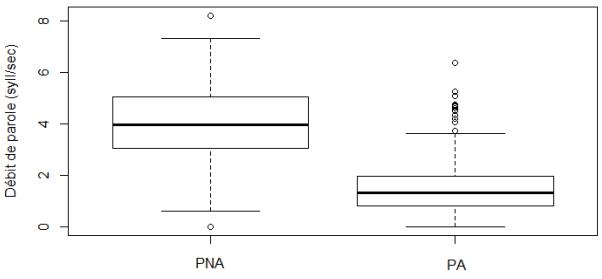



FIGURE 1 – Débit de parole moyen des PNA et des PA dans les groupes de souffle

Si la durée des pauses remplies n'est pas non plus significativement différente entre les deux groupes

($\beta$ = -65.4, SE = 32.4, $p$ = 0.09), avec une durée moyenne de 584.8 ms pour les pauses remplies des PA contre 513.4 ms pour les PNA, c'est en revanche la durée des pauses silencieuses au sein des groupes de souffle qui est significativement différente entre les deux populations ($\beta$ = -290.5, SE = 83.5, $p$ = 0.01), avec une durée moyenne des pauses silencieuses de 682.4 ms pour les PA contre 371.7 ms pour les PNA.

En ce qui concerne la densité gestuelle (nombre de gestes par seconde tous types de gestes confondus), on n'observe pas de différence significative entre les deux groupes ($\beta$ = 0.16, SE = 0.07, $p$ = 0.06) avec une densité gestuelle moyenne de 0.38 gestes/sec pour les PA contre 0.24 gestes/sec pour les PNA. La différence entre les deux groupes reste valable si l'on prend en compte le nombre de gestes en fonction du temps d'articulation ($\beta$ = 0.35, SE = 0.06, $p < 0.005$). Cette absence de significativité pour la densité gestuelle s'applique de la même manière aux gestes non-référentiels ($\beta$ = 0.102, SE = 0.05, $p$ = 0.10) et référentiels ($\beta$ = 0.06, SE = 0.05, $p$ = 0.24). Sur le nombre total de gestes produits, le rapport geste/mot est donc de 0.6 pour les PA contre 0.1 pour les PNA. Si l'on ne compte pas les amorces de gestes et les gestes liés à la recherche lexicale, la proportion reste la même entre les deux groupes avec 0.5 geste/mot pour les PA contre 0.09 geste/mot pour les PNA. Il n'y a pas non plus de différence significative sur la durée des gestes entre les PA et les PNA ($\beta$ = -98.7, SE = 188.4, $p$ = 0.61), que ce soit pour les gestes référentiels ($\beta$ = 433, SE = 342, $p$ = 0.25) ou les gestes non-référentiels ($\beta$ = -363.1, SE = 227.8, $p$ = 0.16).

Les principales différences entre les PA et les PNA se situent dans la structure interne des gestes, dans certaines phases gestuelles. En ce qui concerne la *préparation*, la différence de durée de cette phase n'est pas significativement différente entre les deux groupes ($\beta$ = 120.8, SE = 54, $p$ = 0.06). En revanche, les PA produisent significativement plus de préparations que les PNA ($\beta$ = 0.7, SE = 0.23, $p$ = 0.01). Pour ce qui est de la *tenue pré-réalisation*, les PA ne produisent pas plus de tenues que les PNA ($\beta$ = 1.4, SE = 0.6, $p$ = 0.08), mais lorsqu'ils en produisent, ces phases sont significativement plus longues que pour les PNA ($\beta$ = 79.8, SE = 30.4, $p$ = 0.03). Les phases de *réalisation* ne sont pas d'une durée significativement différente entre les deux groupes ($\beta$ = -12.4, SE = 58.3, $p$ = 0.83), mais elles sont moins nombreuses chez les PA que chez les PNA ($\beta$ = -2.9, SE = 6, $p$ = 0.02). Les phases de *tenues post-réalisation* ne sont ni plus nombreuses dans un groupe que dans un autre ($\beta$ = -0.07, SE = 0.32, $p$ = 0.8), ni plus longues ($\beta$ = -243.7, SE = 170, $p$ = 0.2). La même observation peut être faite pour les *rétractions* qui ne sont ni plus nombreuses dans un groupe que dans l'autre ($\beta$ = -0.29, SE = 0.42, $p$ = 0.5), ni plus longues ($\beta$ = -49.7, SE = 54.9, $p$ = 0.4).

L'ensemble de ces résultats est représenté dans la Figure 2 qui présente la durée moyenne des phases gestuelles et leur pourcentage en fonction du nombre total de gestes produit par chaque groupe. L'absence de significativité sur la durée des tenues post-réalisation s'explique sans doute par la grande variabilité de ces tenues chez les PA, alors que l'absence de significativité du nombre de pré-réalisation entre les deux groupes s'explique par le très petit nombre de ce type de phase chez les PNA, ce qui est en soi révélateur.

# 6 Discussion et conclusion

Cette étude portait sur les liens entre parole et gestualité chez des personnes aphasiques non fluentes (N = 4) en comparaison avec des personnes non-aphasiques (N = 4), à partir d'un corpus

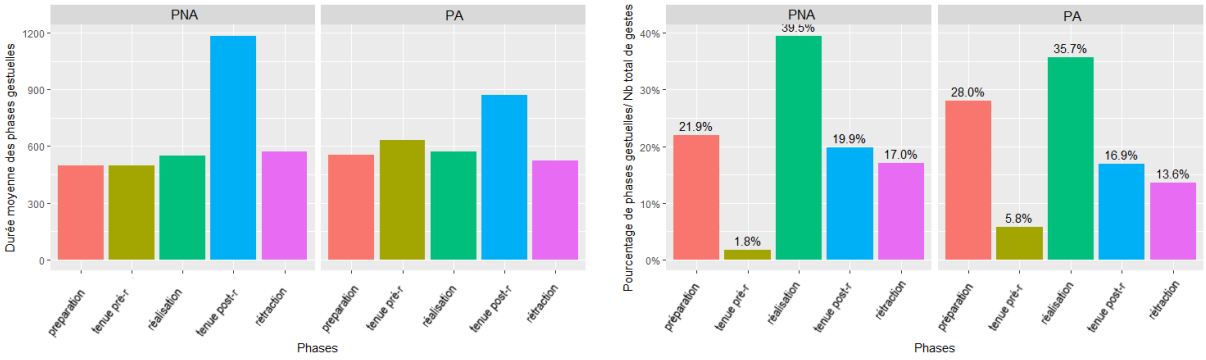



FIGURE 2 – Durée moyennes des phases gestuelles (gauche) et pourcentages d'occurrence des phases gestuelles en fonction du nombre total de gestes (droite) chez les PNA et les PA

d'enregistrements vidéos fourni par AphasiaBank (MacWhinney *et al.*, 2011). L'étude portait principalement sur des récits personnels, e.g. des situations d'interaction où les PA sont certes contraintes par les questions posées par le chercheur, mais dans lesquelles la parole ne peut s'appuyer sur des images présentes dans le contexte d'interaction, ni sur l'iconicité du discours procédural.

Comme l'on pouvait s'y attendre, les résultats montrent que le débit de parole des PA et leur vitesse d'articulation sont beaucoup plus lents que celui des PNA, et que ceci est dû non seulement à des difficultés d'articulation des mots et de nombreuses amorces au niveau verbal, mais aussi à la présence de pauses silencieuses plus longues que chez les PNA, mais pas nécessairement plus nombreuses. Les pauses remplies ne sont pas significativement différentes dans les deux groupes de participants. Comme l'a montré Swerts (1998), les pauses remplies jouent essentiellement un rôle dans la structuration du discours contrairement aux pauses silencieuses qui permettent de gérer des difficultés plus locales et sont en phase avec les nombreuses recherches lexicales que peut engendrer l'aphasie.

En ce qui concerne les gestes, la densité gestuelle (nb de gestes par seconde) n'est pas différente dans les deux groupes de locuteurs, que ce soit pour les gestes référentiels et les gestes non-référentiels. Mais bien entendu, comme le débit de parole des PA est lent que celui des PNA, cela signifie que les PA produisent plus de gestes par nombre de mots que les PNA ce qui confirme les études de Feyereisen (1983) et Cocks *et al.* (2013). Ces résultats sont également compatibles avec ceux des travaux qui trouvent que les PA gestualisent plus que les PNA, et qui sont basés principalement sur le rapport des gestes au nombre de mots. Dans une recherche à venir, ces résultats seront affinés par une étude des types de gestes au sein des catégories un peu larges que sont les gestes référentiels et non-référentiels, avec des pistes prometteuses concernant l'utilisation des battements et des gestes conventionnels par les PA dans ce type d'interaction.

Enfin, on observe des différences de structure interne des gestes entre les deux groupes de participants. Les PA produisent proportionnellement moins de phases de réalisation que les PNA et ceci montre que leur gestualité est liée à leurs difficultés en parole : de même que les PA produisent beaucoup plus d'amorces de mots que les PNA, ils ont plus d'amorces de gestes. Il leur arrive de préparer un geste puis de rétracter les mains sans que ce geste ait été réalisé. On peut donc dire que les abandons gestuels vont de pair avec les amorces verbales. Les PA produisent aussi plus de phases

de préparation que les PNA et là encore, ce schéma correspond à leur production verbale, certains PA s'exprimant principalement sous la forme de groupes nominaux sans liens syntaxiques entre les groupes, leurs gestes sont moins enchaînés. Enfin, la durée plus longue de leurs tenues pré-réalisation montre également que les PA font des pauses plus longues avant de réaliser leurs gestes jusqu'à ce que soit résolues les recherches lexicales. Ceci correspond à ce qui avait déjà été observé dans le bégaiement par Mayberry & Jaques (2000), à savoir que le geste est suspendu – plutôt qu'abandonné – jusqu'à ce que le mot puisse être produit quand la syllabe initiale est plus longue. Il serait d'ailleurs intéressant de voir sur un plus grand nombre de locuteurs si la structure des gestes est identique dans l'aphasie de Broca et l'aphasie transcorticale motrice.

# Remerciements

Je tiens à remercier B. MacWhinney et D. Fromm pour m'avoir donné accès à AphasiaBank, ainsi que tous les locuteurs qui ont participé à cette base de données. AphasiaBank fournit des informations précieuses sur la parole aphasique en mettant à la disposition des chercheurs des enregistrements multimodaux par ailleurs difficiles à collecter. Je remercie aussi deux relecteurs anonymes pour leurs commentaires constructifs sur une version précédente de cet article.

# Références

AHLSÉN E. (1991). Body Communication as Compensation for Speech in a Wernicke's Aphasic—a Longitudinal Study. *Journal of Communication Disorders*, **24**, 1–12.

AHLSÉN E. (2015). Gestures Used in Word Search Episodes – by Persons with and without Aphasia. In K. JOKINEN & M. VELS, Eds., *The 2nd European and the 5th Nordic Symposium on Multimodal Communication*, p. 9–15, Tartu, Estonie.

BATES D., ET AL. (2014). Linear mixed-effects models using eigen and s4. Computer program : http ://cran.r-project.org.

BEEKE S., ET AL. (2015). Conversation focused aphasia therapy : investigating the adoption of strategies by people with agrammatism. *Aphasiology*, **29(3)**, 355–377.

BLOM JOHANSSON M. (2012). *Aphasia and Communication in Everyday Life : Experiences of persons with aphasia, significant others, and speech-language pathologists*. Phd, Department of Public Health and Caring Sciences, Uppsala.

BOERSMA P. & WEENINK D. (2009). Praat : doing phonetics by computer (Version 5.1.05) Computer program : http ://www.fon.hum.uva.nl/praat/.

CICONE M., ET AL. (1979). The Relation between Gesture and Language in Aphasic Communication. *Brain and Language*, **8**, 324–349.

COCKS N., ET AL. (2013). The impact of impaired semantic knowledge on spontaneous iconic gesture production. *Aphasiology*, **27(9)**, 1050–1069.

DANLY M. & SHAPIRO B. (1982). Speech Prosody in Broca's Aphasia. *Brain and Language*, **16**, 171–190.

DE BEER C., ET AL. (2017). How Much Information Do People With Aphasia Convey via Gesture ? *American Journal of Speech-Language Pathology*, **26**, 483–497.

DE RUITER J. P. (2006). Can gesticulation help aphasic people speak, or rather, communicate ? *International Journal of Speech-Language Pathology*, **8(2)**, 124–127.

DIPPER L., ET AL. (2015). The language–gesture connection : Evidence from aphasia. *Clinical Linguistics & Phonetics*, **29(8-10)**, 748–763.

FEYEREISEN P. (1983). Manual Activity During Speaking in Aphasic Subjects. *International Journal of Psychology*, **18**, 545–556.

HOGREFE K., ET AL. (2013). Gestural expression in narrations of aphasic speakers : redundant or complementary to the spoken expression ? In *TIGER*, p. 1–4, Lund, Suède.

KENDON A. (2004). *Gesture. Visible Action as Utterance*. Cambridge : Cambridge University Press.

KROENKE K.-M., ET AL. (2013). Lexical learning in mild aphasia : Gesture benefit depends on patholinguistic profile and lesion pattern. *Cortex*, **49(10)**, 2637–2649.

KUROWSKI K. & BLUMSTEIN S. E. (2016). Phonetic basis of phonemic paraphasias in aphasia : Evidence for cascading activation. *Cortex*, **75**, 193–203.

LOUIS M. (2003). *Etude longitudinale de la dysprosodie d'un cas d'Aphasie Progressive Primaire : analyse des variables temporelles*. Thèse de doctorat, Université d'Aix en Provence.

MACAULEY B. L. & HANDLEY C. L. (2005). Gestures Produced by Patient With Aphasia and Ideomotor Apraxia. *Contemporary Issues in Communication Science and Disorders*, **32**, 30–37.

MACWHINNEY B., ET AL. (2011). AphasiaBank : Methods for Studying Discourse. *Aphasiology*, **25(11)**, 1286–1307.

MAYBERRY R. I. & JAQUES J. (2000). Gesture production during stuttered speech : insights into the nature of gesture-speech integration. In D. MCNEILL, Ed., *Language and Gesture*, Language, Culture & Cognition, p. 199–214. Cambridge : Cambridge University Press.

MOL L., ET AL. (2013). Gesturing by speakers with aphasia : how does it compare ? *Journal of Speech and Hearing Research*, **56(4)**, 1224–1236.

NYSTRÖM M. (2006). Aphasia – an existential loneliness : A study on the loss of the world of symbols. *International Journal of Qualitative Studies on Health and Well-being*, **1(1)**, 38–49.

PREISIG B. C., ET AL. (2015). Perception of co-speech gestures in aphasic patients : A visual exploration study during the observation of dyadic conversations. *Cortex*, **64**, 157–168.

PRITCHARD M., ET AL. (2015). Language and iconic gesture use in procedural discourse by speakers with aphasia. *Aphasiology*, **29(7)**, 826–844.

R CORE TEAM (2012). A language and environment for statistical computing. R foundation for statistical computing. Computer program : http ://www.r-project.org.

ROSE M. L., ET AL. (2015). Comparing multi-modality and constraint-induced treatment for aphasia : a preliminary investigation of generalisation to discourse. *Aphasiology*, **30(6)**, 678–698.

SLOETJES H. & WITTENBURG P. (2008). Annotation by category – ELAN and ISO DCR. In *6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Maroc.

SWERTS M. (1998). Filled Pauses as Markers of Discourse Structure. *Journal of Pragmatics*, **30**, 485–496.

TULLER B. (1984). On Categorizing Aphasic Speech Errors. *Neuropsychologia*, **22(5)**, 547–557.

# « Tout ça c'est abstrait » Comment le degré d'abstraction d'un mot expliqué affecte-t-il la parole multimodale ?

Marion Tellier[1], Gale Stam[2], Alain Ghio[1],

(1) Aix-Marseille Univ, CNRS, LPL, UMR 7309, Aix-en-Provence, France
(2) National Louis University, Department of Psychology, Skokie, IL USA

marion.tellier@lpl-aix.fr; gstam@nlu.edu; alain.ghio@lpl-aix.fr

## RÉSUMÉ

Lorsque nous parlons, nous produisons spontanément des gestes coverbaux, c'est-à-dire en lien avec la parole. Dans cette étude nous nous intéressons à des séquences orales d'explication lexicale de mots concrets et de mots abstraits en nous demandant si le degré d'abstraction du sujet de l'interaction affecte la production gestuelle. Dans la mesure où les mots concrets sont liés à une grande iconicité (*high imageability*) et les mots abstraits à une faible iconicité (*low imageability*) (Paivio, 1986 ; Palmer *et al.,* 2013), on peut faire l'hypothèse que les gestes qui accompagnent les mots concrets sont majoritairement iconiques et que les gestes qui accompagnent les mots abstraits sont principalement métaphoriques (si on se base sur cette distinction de McNeill, 1992). Nous examinerons plusieurs paramètres: le taux gestuel du discours, le type de geste employé et l'espace gestuel utilisé.

## ABSTRACT

When we speak, we spontaneously produce cospeech gestures. In this study we are interested in oral sequences of lexical explanation of concrete words and abstract words. We are questioning whether the degree of abstraction of the topic of the interaction affects gestural production. Insofar as concrete words are linked to high imageability and abstract words to low imageability (Paivio, 1986 ; Palmer *et al.*, 2013), we can assume that the gestures that accompany concrete words are mostly iconic and the gestures produced with abstract words are mainly metaphorical (based on McNeill's distinction, 1992). We will examine several parameters: the gesture rate, the type of gesture used and the use of gesture space.

MOTS-CLÉS : gestes coverbaux, concret, abstrait, explication lexicale

KEYWORDS: cospeech gestures, concrete, abstract, lexical explanation

Lorsque nous parlons, nous produisons spontanément des gestes coverbaux, c'est-à-dire en lien avec la parole (McNeill, 1992; 2005 ; Kendon, 2004). L'objectif général de cette étude est d'analyser, dans une interaction, les processus d'adaptation multimodale (voco-verbale et gestuelle) d'un locuteur en fonction du contenu du message et de son interlocuteur. En ce qui concerne la variation de la nature du contenu du message, nous avons choisi de faire varier l'aspect abstrait *vs* concret. En effet, dans cette étude nous nous intéressons à des séquences d'explication lexicale de mots concrets et de mots abstraits en nous demandant si le degré d'abstraction du sujet de l'interaction affecte la production gestuelle. En ce qui concerne la variation de la nature de l'interlocuteur, nous avons


10.21437/JEP.2018-38

choisi de faire varier la maîtrise de la langue de l'interaction par l'interlocuteur. Ainsi, les destinataires de l'explication lexicale, sont tour à tour des locuteurs natifs de la langue française ou des apprenants de français langue étrangère. Ces processus sont étudiés à la fois sur des aspects oraux (au sens de voco-verbal) mais aussi gestuels car nous considérons que ces deux canaux appartiennent à un seul et même système de production (voir entre autres McNeill, 1992, 2005). Notre hypothèse est que l'explication de mots abstraits élicite une gestuelle différente de l'explication de mots concrets, notamment en termes d'iconicité du geste et d'utilisation de l'espace gestuel.

# 1 Le monde abstrait et le monde concret

Les mots concrets sont liés à une grande iconicité notamment en termes de représentation mentale alors que les mots abstraits sont plutôt encodés verbalement (Paivio, 1986). Les mots concrets sont davantage associés à des informations contextuelles et à des expériences sensori-motrices que les mots abstraits (Schwanenflugel, Harnishfeger, & Stowe, 1988 ; Marques & Nunes, 2012 ; Palmer *et al.* 2013). En ce qui concerne les gestes coverbaux, ils sont caractérisés par leur production spontanée, en d'autres termes, ils sont uniques et personnels et ne font pas partie d'un répertoire fixe (McNeill, 1992). Les travaux de McNeill (1992) et McNeill et Duncan (2000) montrent que geste et parole forment un tout et sont liés à la pensée et que l'analyse des gestes révèle les représentations mentales de l'individu activées pendant la production langagière. Le geste et la parole reflètent la même idée sous-jacente, au même moment, mais n'en expriment pas nécessairement les mêmes aspects (le geste peut véhiculer des informations complétant celles apportées par la parole) (Goldin-Meadow, 2003). Nous nous focaliserons ici sur deux dimensions gestuelles établies par McNeill (1992, 2005). Tout d'abord, les « iconiques » (« iconic gestures ») qui entretiennent une relation très étroite et visuelle avec le contenu sémantique du référent (1992 : 78), par exemple lorsque l'on montre la taille ou la forme d'un objet décrit avec les mains ou lorsque l'on mime l'action de tenir un volant en disant « conduire ». Ensuite, les « métaphoriques » (« metaphoric gestures») qui représentent des concepts abstraits et des métaphores. « Metaphoric gestures create images of abstractions. In such gestures, abstract content is given form in the imagery of objects, space, movement, and the like» (McNeill 1992: 145). Les exemples les plus fréquents sont ceux où les mains en forme de contenant (une main ou les deux semblant tenir un bol) sont utilisées pour accompagner un concept abstrait, comme si le concept était tenu dans la main. Le geste métaphorique est tout particulièrement utilisé avec des métaphores verbales (Cienki & Müller, 2008). Depuis quelques années, le concept même de geste métaphorique a été remis en question, notamment en raison de la difficulté à distinguer certains métaphoriques de certains iconiques. Néanmoins, dans le corpus présenté ici, l'annotation des gestes a été réalisée en suivant les dimensions définies par McNeill (1992, 2005).

Dans la mesure où, comme l'ont souligné, entre autres, Pavio (1986) et Palmer *et al*. (2013), les mots concrets sont liés à une grande iconicité (*high imageability*) et les mots abstraits à une faible iconicité (*low imageability*), on peut faire l'hypothèse que les gestes qui accompagnent les mots concrets sont majoritairement iconiques et que les gestes qui accompagnent les mots abstraits sont principalement métaphoriques (si on se base sur cette distinction de McNeill, 1992). À notre connaissance, aucune étude n'a comparé les gestes produits lors d'explication orale de concepts concrets et abstraits.

# 2 Méthodologie

## 2.1 Le projet Gesture in Teacher Talk (GTT)

Le projet GTT a commencé en 2009 et est le fruit d'une collaboration franco-américaine. Il a permis de recueillir un corpus audiovisuel de 7h d'interaction dans le cadre d'une tâche d'explication lexicale dans un contexte d'enseignement de Français Langue Etrangère (FLE). La situation interactionnelle est fondée sur un jeu de devinette dans lequel une personne doit faire deviner une série de mots à une autre personne.

### *2.1.1 Participants*

Dix étudiants en première année de Master FLE à l'université d'Aix-Marseille ont participé à l'expérience dans le rôle de la personne qui doit faire deviner les mots (situation d'enseignant). Dans le rôle de la personne qui doit deviner (situation d'apprenant), nous avons sélectionné dix étudiants français, de langue maternelle française, issus d'autres cursus et dix étudiants étrangers apprenants de français (d'un niveau B1/B2 selon le CECRL, 2001). Ce recrutement différentiel est nécessaire pour pouvoir observer les processus d'adaptation de la communication en fonction de l'interlocuteur, notamment sa maitrise du français. Dans l'analyse, ce sont essentiellement les procédés mis en œuvre chez la personne en situation d'enseignant qui nous intéressent.

### *2.1.2 Matériel*

Une douzaine de mots a été sélectionnée : trois verbes, trois noms, trois adjectifs et trois adverbes. Dans chaque catégorie grammaticale, nous avons placé deux concepts concrets et un concept abstrait. Les mots concrets étaient : *grimper, emballer, océan, trottoir, rapé, usé, doucement* et *rapidement*. Les mots abstraits : *jalousie, se souvenir, approximativement* et *fier*. Les mots concrets sont des mots qui évoquent une image ou une action que l'on peut se représenter visuellement. Ces mots ont été imprimés sur des étiquettes (un mot par étiquette) et placés dans une boite pour pouvoir être tirés au sort.


FIGURE 1 : Exemple d'annotation du corpus GTT avec le logiciel *Elan*

### *2.1.3 Procédure*

Chaque étudiant de master FLE (désormais EMF) se voyait attribuer aléatoirement un partenaire natif et un partenaire non natif[1]. Même si les partenaires ne se connaissaient pas, l'EMF était informé au préalable sur le statut de son partenaire (natif ou non natif). L'EMF tirait au sort une étiquette et devait faire deviner le mot inscrit à son partenaire. Les seules contraintes étaient de ne pas utiliser de mots de la même famille ni de mots dans d'autres langues que le français. L'ordre de passage (natif/non natif) était contre-balancé. Cette procédure nous permet d'éliciter des explications des mêmes mots chez différents sujet EMF et avec deux partenaires différents. Nous obtenons alors des productions verbales et gestuelles comparables. Cependant, même si cette tâche est contrôlée, elle n'en demeure pas moins spontanée et interactive puisque les participants se focalisent sur l'enjeu de la tâche : faire deviner les mots. Notre corpus est donc divisé en séquences d'explications (une séquence par EMF, par mot et par condition). Au total, il est composé de 10 EMF qui expliquent 12 mots à deux partenaires soit 240 séquences d'explications. Dans le cadre de cet article, nous nous focaliserons uniquement sur les interactions entre locuteurs natifs en comparant la production gestuelle dans le cadre des explications de mots abstraits *vs* concrets, soit 120 séquences d'explication: 8 mots concrets et 4 mots abstraits.

## 2.2 Annotation du corpus

L'annotation a été réalisée avec le logiciel *Elan* (Sloetjes & Wittenburg, 2008) dont la particularité est de pouvoir transcrire et d'annoter les phénomènes multimodaux de la parole sous forme de partition (une piste par aspect transcrit ou annoté) comme le montre la Figure 1.

### *2.2.1 Les types de gestes*

Nous avons choisi pour cette étude les catégories gestuelles de McNeill (1992) qui sont à considérer davantage comme des dimensions que des types restrictifs (un même geste pouvant avoir plusieurs dimensions) (McNeill, 2005) : les déictiques (pointage), les iconiques, les métaphoriques et les battements (rythmant la parole, sans contenu sémantique). Nous y avons ajouté des catégories complémentaires (voir Tellier & Stam, 2012 pour le détail) : les emblèmes (culturel, conventionnel), les Butterworth (de recherche lexicale) et les interactifs (adressés à l'interlocuteur pour la gestion de l'interaction). Les gestes du corpus ont été annotés par une personne puis vérifiés intégralement et séparément par deux annotateurs gestualistes. Lorsque des désaccords sont intervenus sur le choix des étiquettes d'annotation, une discussion entre les deux gestualistes a abouti à un consensus sur l'étiquetage du geste (la même procédure a été effectuée pour l'annotation de l'espace gestuel). Tous les types de gestes présentés ici ne seront pas analysés en détail dans cette étude.

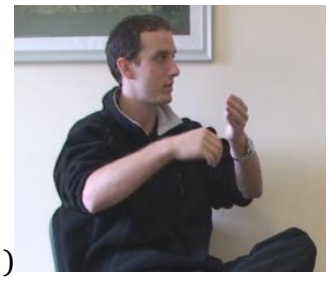
1)  2) 

FIGURE 2 : 1) geste iconique mimant l'action de « raper » et 2) geste métaphorique pour mettre en évidence le concept d'escalader

[1] Nous utilisons les termes de locuteurs natifs et non natifs par commodités bien qu'ils soient controversés (Joseph, 2013).

### *2.2.2 L'espace gestuel*

L'utilisation de l'espace gestuel demeure une composante encore peu étudiée dans les études de la gestuelle car difficile à annoter. Nous avons utilisé le schéma de McNeill (1992) pour l'analyse de l'espace gestuel (Figure 3). Ce diagramme soulève un problème majeur : il est en deux dimensions alors que le geste est produit dans un espace tridimensionnel. Ainsi, un geste produit dans le centre-centre près du corps est codé de la même façon qu'un geste produit dans le centre-centre mais les bras tendus vers l'avant. Or ces deux gestes n'ont clairement pas la même taille et ne sont pas perçus de la même façon par l'interlocuteur. Nous avons donc ajouté la dimension « bras tendu devant » pour annoter les gestes projetés vers l'avant (comme dans la Figure 1). L'intérêt d'annoter l'espace gestuel est qu'il nous renseigne sur la taille du geste et sa visibilité pour l'interlocuteur.

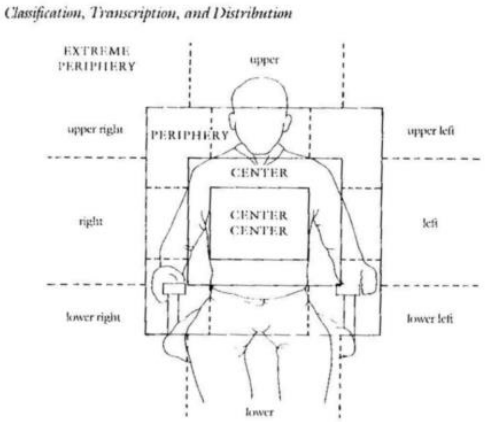



FIGURE 3 : Espace gestuel (McNeill, 1992 : 89)

### *2.2.3 La quantité de geste*

Le taux gestuel est une façon intéressante de comparer la quantité de gestes produits dans des interactions de durée différentes. Il existe plusieurs méthodes pour le calculer (voir Tellier, 2014 pour une comparaison). Ici, nous avons calculé le nombre de gestes par mots (i.e. le taux gestuel).

# 3 Résultats

Dans cet article, nous n'exposerons qu'une partie des résultats qui concerne seulement l'axe de variation concret *vs* abstrait en interaction avec des interlocuteurs natifs. L'objectif est de répondre à la question suivante : quelles différences l'explication de mots concrets et de mots abstraits dans le cadre d'une devinette lexicale engendre-t-elle au niveau de la parole multimodale ?

## 3.1 Analyse statistique

Les traitements statistiques ont été réalisés avec le logiciel R 3.4.2 Dans les analyses de variances, le "type de mot" (concret/abstrait) est le facteur explicatif. Chaque locuteur EMF ayant participé aux deux conditions, nous utilisons des ANOVA à mesure répétée dans laquelle le participant est le facteur de répétition. Pour chaque analyse, nous vérifions la normalité des distributions (test de Shapiro) et l'homogénéité des variances (test de Bartlett).

## 3.2 Au niveau de l'interaction

Est-ce qu'un mot abstrait est plus difficile à expliquer qu'un mot concret ? On peut se baser sur le temps d'explication du mot. Le calcul de cette durée s'arrête lorsque le mot est trouvé par le partenaire.

| | *Durée interaction (s)* | *Temps de parole du locuteur 1 (s)* | *Nombre de mots produits (loc. 1)* | *Temps de gestualisation (s)* | *Nombre de gestes produits (loc.1)* | *Taux gestuel (nombre de gestes/mots)* |
|---|---|---|---|---|---|---|
| Abstrait | 21.88 SD 7,28 | 14.13 SD 4.98 | 42.8 SD 13.41 | 7.23 SD 4.9 | 4.93 SD 3.20 | 0.11 SD 0.06 |
| Concret | 13.8 SD 5.3 | 9.20 SD 3.24 | 28.36 SD 9.44 | 4.07 SD 2.03 | 2.7 SD 1.25 | 0.10 SD 0.05 |
| *p-value* | **0.018 *** | **0.016 *** | **0.017*** | 0.052 . | **0.04 *** | 0.4 . |

TABLE 1 – Moyenne et écart-type des paramètres généraux en fonction du type de mot

L'abstraction a un impact sur la durée de l'interaction et sur les aspects verbaux de façon significative (Table 1). L'explication de mots abstraits a donc pris plus de temps et les interlocuteurs ont parfois eu plus de difficulté à deviner le mot-cible. Si le temps d'explication ainsi que le nombre de mots utilisés sont significativement plus importants avec un mot abstrait, la production gestuelle semble moins nettement affectée. En effet, seul le nombre de gestes produits est légèrement supérieur avec des mots abstraits mais le taux gestuel demeure similaire. On peut voir, de manière plus qualitative comment les explications varient en fonction du degré d'abstraction du mot. Les deux explications ci-dessous ont été produites par la même locutrice. On voit que pour expliquer « approximativement », son discours est marqué par de nombreuses disfluences, des pauses remplies

**Explication du mot concret « trottoir »**

| | | |
|---|---|---|
| 1<br>2 | Loc1 | hum *(889ms)* au niveau de des euh des rues *(237ms)* il y a *(452ms)* le bitume avec les voitures qui circulent *(284ms)* et les piétons qui sont sur le |
| 3 | Loc2 | Trottoir |
| 4 | Loc1 | Ouais |
| *Description* | | Temps total de l'explication: 10.89s / 8.2s de parole pour loc 1 / 26 mots pour loc 1 / 2 gestes produits (2 iconiques) / espace gestuel utilisé : 1 bras tendu et 1 extrême périphérie. |

**Explication du mot abstrait « Approximativement »**

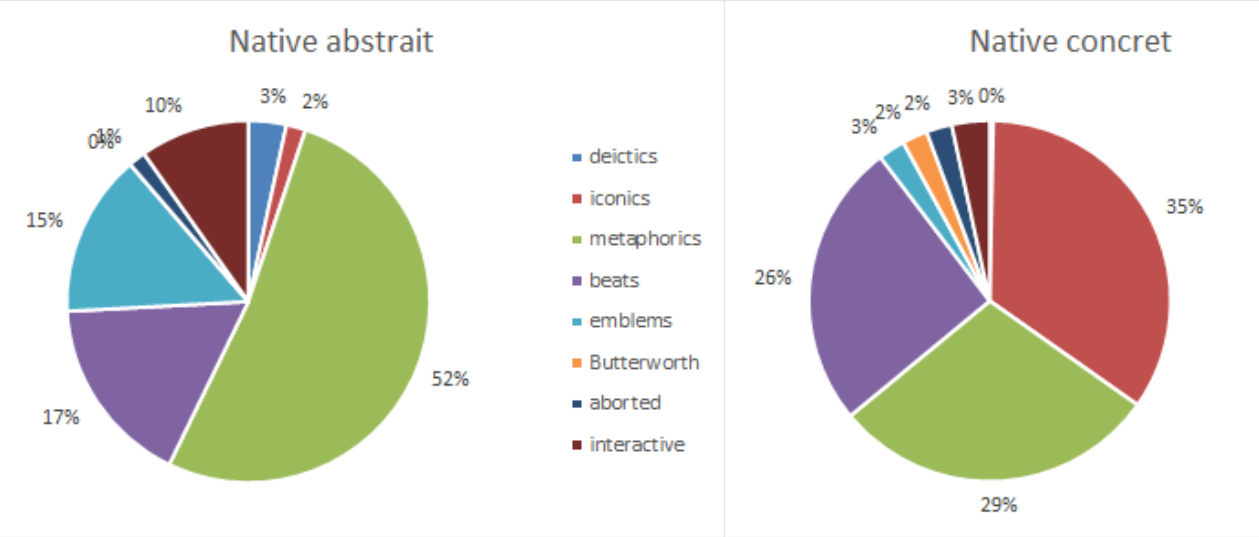
| | | |
|---|---|---|
| 1<br>2<br>3 | Loc1 | alors celui-là *(208ms)* hum *(673ms)* lorsque tu donnes une réponse euh *(250ms)* une réponse euh *(460ms)* euh *(367ms)* qui n'est *(226ms)* qui n'est pas *(686 ms)* c'est on parle *(269ms)* qui qui n'est pas comment dire *(841ms)* euh *(1549ms)* précise *(496ms)* |
| 4 | Loc2 | Hum |
| 5 | Loc1 | on dit que tu réponds |
| 6 | Loc2 | Approximativement |
| 7 | Loc1 | v- voilà |
| *Description* | | Temps total de l'explication: 21.79s / 12.7 s de parole pour loc 1 / 38 mots pour loc 1 / 7 gestes produits (6 métaphoriques et 1 emblème) / espace gestuel utilisé : 5 dans le centre, 1 dans le centre centre et 1 bras tendu. |

De façon globale, l'abstraction nécessite un temps d'explication bien plus important. Qu'en est-il de la nature des gestes ?

## 3.3 Le degré d'abstraction du mot affecte-t-il le type de gestes utilisés ?

### *3.3.1 Dimension principale du geste*

Pour observer l'adaptation gestuelle de l'EMF en fonction de la nature abstraite/concrète des mots à expliquer, nous avons calculé la proportion de type de gestes (déictique, iconique, métaphorique…) en fonction du nombre total de gestes produits (Table 2).




| | Iconiques | Métaphoriques | Emblèmes |
|---|---|---|---|
| Abstrait | 0.01 SD 0.02 | 0.52 SD 0.2 | 0.14 SD 0.18 |
| Concret | 0.31 SD | 0.26 SD 0.15 | 0.02 SD 0.03 |
| *p* | 0.001 ** | 0.03 * | 0.08 . |

TABLE 2– Répartition des types de geste en fonction de la nature abstraite/concrète des mots à expliquer (en % sur les graphiques, en proportion à 1 dans le tableau)

Les gestes iconiques sont essentiellement présents dans les explications de mots concrets (0.31 *vs* 0.01, p=0.001). Les gestes métaphoriques sont plus présents dans les explications de mots abstraits (0.52 *vs* 0.26, p=0.03), ainsi que les emblèmes (0.14 *vs* 0.02) mais dans ce cas, le résultat est une tendance non significative (p=0.08). Ceci peut être expliqué par le fait que les emblèmes sont des métaphores codifiées (comme des expressions idiomatiques gestuelles). Par ailleurs, l'explication du mot "approximativement" a généré une production importante de l'emblème qui est souvent associé à cet adverbe (basculement de droite à gauche de la main à plat).

### *3.3.2 Utilisation de l'espace gestuel*

Pour observer l'adaptation gestuelle spatiale de l'EMF en fonction de la nature abstraite/concrète des mots à expliquer, nous avons calculé la proportion de type de gestes dans les zones centrales *vs* périphériques en fonction du nombre total de gestes produits (Table 3). Pour l'analyse statistique, nous avons regroupé les 2 zones centrales (center_center + center) *vs* les 3 zones périphériques. Nous constatons que l'explication de concepts abstraits entraine une centralisation de l'espace gestuel.

| | Central | Périphérique |
|---|---|---|
| Abstrait | 0.77 SD 0.19 | 0.22 SD 0.19 |
| Concret | 0.57 SD 0.29 | 0.42 SD 0.29 |
| *p* | 0.04* | 0.04 * |

TABLE 3– Répartition de l'espace gestuel en fonction de la nature abstraite/concrète des mots à expliquer

# 4 Conclusion

Notre étude montre que l'explication de concepts abstraits donne lieu à des explications significativement plus longues et à un nombre de geste produits plus important (bien que le taux

gestuel ne soit pas affecté) que l'explication de mots concrets. En outre, plus le degré d'abstraction des mots expliqués est élevé, plus la gestuelle concomitante est métaphorique. Similairement, l'explication de mots concrets élicite une gestuelle plus iconique (en lien avec les représentations mentales du locuteur). On peut supposer que le mot abstrait est plus difficile à deviner pour l'interlocuteur à la fois parce que l'explication est plus disfluente (plus de pauses et de reprises) mais également parce qu'elle est accompagnée de moins d'illustrations, qu'elles soient verbales ou gestuelles. Par ailleurs, l'analyse de l'espace gestuel nous apprend que lors de l'explication de mots abstraits, les locuteurs utilisent un espace plus restreint, alors que l'explication de mots concrets entraine des gestes plus larges et donc plus visibles. Cela peut être lié à l'iconicité des gestes (des actions, des formes, des localisations d'éléments…) qui nécessite un plus grand espace tandis que les gestes de l'abstrait peuvent être restreints au centre de l'espace. Rares sont les études à notre connaissance qui ont analysé l'utilisation de l'espace en fonction du type de geste produit et du topic conversationnel. McNeill (1992) a montré que les gestes métaphoriques avaient tendance à apparaitre dans la partie « lower center space » tandis que les iconiques étaient plutôt produits dans les régions « center-center » et « center » (p. 88-90) mais cela ne recouvre pas entièrement nos résultats. Cependant, McNeill n'a étudié que des narrations et non des explications lexicales, la différence de tâche langagière ainsi que son enjeu interactif (expliquer suffisamment clairement le mot pour que son partenaire le devine) peut avoir un effet sur l'utilisation de l'espace gestuel et la visibilité du geste. Il serait à présent pertinent d'examiner si les mêmes tendances se retrouvent dans l'explication de mots concrets et abstraits à des partenaires non natifs afin de vérifier si la présence d'un interlocuteur apprenant élicite des gestes plus iconiques et plus grands (afin de faciliter la compréhension de l'interlocuteur), y compris lors de l'explication de concepts abstraits.

# Références


CIENKI, A and CORNELIA M (2008). Metaphor, gesture and thought. In: Raymond W. Gibbs (ed.), *Cambridge Handbook of Metaphor and Thought*, 483–501. Cambridge: Cambridge University Press.

CONSEIL DE L'EUROPE (2001). *Un cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. Strasbourg. http://www.coe.int/t/dg4/linguistic/cadre_FR.asp

GOLDIN-MEADOW, S. (2003). *Hearing gesture: how our hands help us think*. Belknap Press of Harvard University Press.

JOSEPH J. (2013). Le corps du locuteur natif : discipline, habitus, identité. *Histoire Épistémologie Langage*, 35(2), 29-45

KENDON, A. (2004). *Gesture. Visible action as utterance*. Cambridge : Cambridge University Press.

MARQUES, J. F., & NUNES, L. D. (2012). The contributions of language and experience to the representation of abstract and concrete words: Different weights but similar organizations. Memory & Cognition, 40(8), 1266-1275.

MCNEILL, D. & DUNCAN, S. (2000). Growth points in thinking-for-speaking. In Mcneill, D. (ed.) *Language and Gesture,* (pp. 141-161). Cambridge : Cambridge University Press.

MCNEILL, D. (1992). *Hand and Mind : What gestures reveal about thought*. Chicago : The University of Chicago Press.

MCNEILL, D. (2005). *Gesture & thought*. Chicago : The University of Chicago Press.

PAIVIO, A (1986). Mental representations: a dual coding approach. Oxford. England: Oxford University Press.

PALMER, S. D., MACGREGOR, L. J., & HAVELKA, J. (2013). Concreteness effects in single-meaning, multi-meaning and newly acquired words. *Brain Research*, 1538, 135-150.

SCHWANENFLUGEL, P. J., HARNISHFEGER, K. K., & STOWE, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *JML*, 27(5), 499-520.

SLOETJES H., WITTENBURG P. (2008). Annotation by category – ELAN and ISO DCR. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008*), Marrakech, Maroc, 816-820.

TELLIER, M. & STAM, G. (2012). Stratégies verbales et gestuelles dans l'explication lexicale d'un verbe d'action, in Rivière, V. (2012) *Spécificités et diversité des interactions didactiques* (pp. 357-374). Paris : Riveneuve éditions. ISBN : 978-2-36013-093-1.

TELLIER, M. (2015). Quelques orientations méthodologiques pour étudier la gestuelle dans des corpus spontanés et semi-contrôlés. *Discours*, 15. [en ligne] http://discours.revues.org/8917

# Effet de la situation de parole sur la variabilité des voyelles en français

Mélanie Lancien[1,2], Nicolas Audibert[1], Cécile Fougeron[1]

(1) Laboratoire de Phonétique et Phonologie, 19 rue des bernardins, 75005 Paris, France
(2) Section des Sciences du Langage et de l'Information,
Quartier Unil-Chamberonne, CH-1015 Lausanne, Suisse
melanie.lancien@unil.ch, nicolas.audibert@sorbonne-nouvelle.fr,
cecile.fougeron@sorbonne-nouvelle.fr

## RESUME

Cette étude s'intéresse à la variation dans la réalisation des voyelles en fonction de la situation de production de la parole. Huit locutrices francophones ont participé à un protocole induisant des interactions naturelles à travers un jeu contrôlé. Cette situation de jeu (J) est comparée à la lecture rapide (LR), la lecture normale (LN), et la lecture pour un malentendant (ME) des mêmes mots cibles. Nous étudions la durée, la dispersion dans l'espace F1*F2, et la variabilité intra-catégorie sur le plan F1*F2*F3 de 1702 exemplaires des voyelles /i, y, E, e, a, O, o, u/. La situation de production affecte toutes les variables. Comparé à LN, les voyelles en J et LR sont plus courtes, elles sont aussi plus centralisées et moins variables en LR, mais pas en J. La comparaison des tailles d'effet montre que la condition de production influence plus fortement la variabilité intra-catégorie qu'elle n'influence la centralisation du système.

## ABSTRACT

**Effect of the speech situation on the variability of vowels in French.**

This study focuses on the variation of French vowels according to speech production situations. Eight female speakers participated in a protocol inducing natural interactions in a controlled game context. This game condition (J) is compared to a fast reading (LR), a normal reading (LN), and reading for a hard of hearing listener (ME) of the same target words. 1702 token of vowels /i, y, E, e, a, O, o, u/ were analyzed. We analyze their duration, the dispersion/centralization in the F1*F2 space, and the intra-category variability on the F1*F2*F3 space. The production situation affects all variables. Compared to LN, vowels in J and LR are shorter, they are also more centralized and less variable in LR, but not in J. The comparison of effect sizes also shows that overall, the production situation has a larger effect on intra-category variability than on the centralization of the system.

MOTS-CLES : styles de parole, espace vocalique, français, parole conversationnelle

KEYWORDS: speech styles, vocalic space, French, conversational speech



10.21437/JEP.2018-39

# 1 Introduction

La parole est modulée par plusieurs sources de variation linguistique et extra-linguistique telles que le contexte segmental, la prosodie, l'état de santé ou les émotions du locuteur, la situation de communication ou la tâche de production. Les études pionnières sur ces deux derniers facteurs, communément regroupés comme des différences de « styles de parole », se sont concentrées sur la comparaison entre parole lue et parole spontanée. Plus tard, des études comme celle de Harmegnies et al. (1994), ont tenté de mettre en évidence les « sous-styles » de « discours spontané » en français. Ils ont comparé les productions d'un même locuteur dans six situations de communication différentes : une tâche de lecture, deux tâches de description sans interlocuteur, une tâche de description interactive dans laquelle le locuteur devait expliquer une image à un auditeur qui pouvait demander des détails, un monologue spontané (interview guidée) et une conversation avec l'expérimentateur. Les auteurs ont remarqué que les propriétés acoustiques de la parole étaient différentes pour chaque tâche et pouvaient former des « familles de style » homogènes. Par exemple, le système vocalique (sur un plan F1*F2) apparaît plus centralisé dans les tâches de description que dans la parole spontanée, les F1 des voyelles sont aussi plus bas dans les tâches de description par rapport à la lecture. Cependant, rien n'indiquait si l'interactivité de la tâche avait un effet sur les paramètres acoustiques de la parole lors des tâches de description. Sur ce sujet Nakamura et al. (2008) ont rapporté en japonais que les voyelles étaient spectralement plus réduites une situation interactive que dans un monologue spontané. Cependant dans cette étude les deux styles n'étaient pas produits par les mêmes locuteurs. La même réserve peut être formulée concernant la plupart des études sur les grands corpus (voir par exemple Audibert et al., 2015) où styles et locuteurs co-varient. Dans ces études, il n'est donc pas possible de distinguer la variation due au style de la variation inter-locuteur. Un des objectifs de la présente étude sera donc d'étudier les effets de la situation de production de la parole pour les mêmes locuteurs.

Un second objectif de cette étude est de mieux comprendre la nature des variations phonétiques entre les situations de production. Pour cela, il nous semble qu'une description multidimensionnelle sera plus à même de capturer la diversité des changements acoustiques possibles. Jusque-là, les variations phonétiques en fonction du style ont été documentées principalement pour des voyelles. L'un des paramètres les plus fréquemment utilisés pour mesurer les distorsions de l'espace vocalique est un indice de centralisation. Cet indice s'exprime comme la distance de chaque exemplaire, ou catégorie vocalique, au centroïde du système dans un espace F1*F2. Toutefois, plusieurs auteurs ont souligné qu'une combinaison de différentes métriques est plus pertinente pour caractériser les différentes façons dont les systèmes et les catégories vocaliques peuvent varier par rapport à la cible (Ferguson et Kewley-Port, 2007 ; Harmegnies et Poch-Olivé, 1992). Il est ainsi possible d'observer d'une part la variation dans la réalisation des cibles acoustiques à travers la dispersion des exemplaires au sein de leur catégorie vocalique, mais aussi des modifications de l'espace des contrastes acoustiques, en termes de centralisation des voyelles vers une voyelle centrale neutre et/ou en termes de chevauchement ou perte de contraste acoustique entre les catégories vocaliques (Fougeron et Audibert, 2011 ; Audibert et al., 2015).

Selon le modèle H & H de (Lindblom, 1990), les locuteurs adaptent leur prononciation afin de satisfaire les contraintes biomécaniques et linguistiques avec un effort minimal, mais aussi selon les informations nécessaires à l'auditeur : on hyper-articule pour maximiser l'intelligibilité quand cela s'avère nécessaire, autrement on hypo-articule. Cependant, ce qui rend le discours « maximalement intelligible » ou « clair » n'est pas si simple. Le « clear speech » semble plutôt se situer dans un continuum que relever d'une catégorie bien définie, comme suggéré par Scarborough et Zellou (2013). Dans ce continuum hypo-hyper, la parole produite dans un discours interactif pourrait

combiner certaines des propriétés du discours hypo-articulé (par exemple, la centralisation, le raccourcissement vocalique) du fait des contraintes temporelles lié à l'interaction ou de l'apport d'informations par le contexte (permettant moins de précision articulatoire), ainsi que des caractéristiques du « clear speech » de façon à satisfaire le besoin d'intercompréhension entre les locuteurs. Par conséquent, le discours spontané pourrait être vu comme un phonogenre contenant des phonostyles ayant différentes propriétés, parmi lesquels se trouverait la parole interactive. Ainsi, Mathon (2014) a montré l'existence d'un phonogenre « commentaire sportif en direct » incluant plusieurs phonostyles selon le type d'événement, et Hupin et Simon (2009) ont mis en évidence différents styles de discours radio selon le type de station de radio.

Dans cette étude, nous évaluerons les variations acoustiques, et plus particulièrement formantiques, dans la parole d'un ensemble de locutrices du français produisant le même matériel dans différentes situations de production de la parole. Un intérêt particulier est porté à la production des voyelles dans un véritable jeu interactif par rapport aux conditions de contrôle lues et simulées. Ce premier travail est destiné à poser les bases d'un projet plus vaste qui visera à établir une typologie de discours spontanés, en prenant particulièrement en compte la présence et l'identité de l'interlocuteur.

# 2 Méthode

## 2.1 Les conditions de production

Afin d'explorer les propriétés de la parole produite dans un cadre véritablement interactif, les mêmes locutrices ont été placées dans quatre conditions de parole dans lesquelles elles ont produit le même matériel phonétique. Chaque enregistrement a été réalisé dans une pièce insonorisée avec des micros AKGC520. Les conditions de production sont les suivantes :

(a) la tâche de jeu (J) : un jeu de cartes a été créé pour combiner la spontanéité de l'expression et le contrôle expérimental des enregistrements et du contenu linguistique. Quatre dyades d'amies ont joué chacune leur tour, l'expérimentateur étant présent uniquement pour le décompte des points. La première joueuse (P1) devait utiliser différents indices pour aider la seconde joueuse (P2) à deviner un mot mystère. Sur chaque carte, un mot mystère (par exemple « livre ») et deux indices (par exemple « auteur », et « lecteur »), étaient présentés à P1. L'utilisation de ces indices pour faire deviner le mot cible rapportait davantage de point à P1, ce qui nous a permis d'éliciter indirectement leur production, Ces mots indices étaient les mots cibles de notre expérience. P1 et P2 changeaient de rôle toutes les 5 cartes, permettant l'enregistrement d'une quantité similaire de données par locuteur dans le même jeu. Chaque joueur avait 51 cartes (distribuées dans le même ordre pour chaque partie) et pouvait ainsi produire jusqu'à 204 mots (indices) cibles. Pour susciter une plus grande vivacité des échanges, une contrainte temporelle a été introduite.

Compte tenu de la durée de la session de jeu (2h en incluant les explications et l'entraînement), l'enregistrement des autres conditions de production a eu lieu 2 à 4 jours plus tard. Ces autres conditions consistaient en trois types de lecture de la liste de mots-cibles introduits dans le jeu :

(b) la lecture normale (LN) dans laquelle les locuteurs sont conviés à lire la liste au rythme qui leur convient

(c) la lecture rapide (LR) dans laquelle les locuteurs sont conviés à lire la liste « aussi vite que possible »

(d) la lecture « pour malentendant » (ME) dans laquelle les locuteurs sont conviés à lire « comme si ils parlaient à leur vieille grand-mère malentendante », de façon à simuler une parole 'claire'

## 2.2 Locutrices

Huit locutrices francophones natives âgées de 18 à 40 ans (moyenne = 23,75) ont participé à l'expérience. Le jeu se jouant en binôme, nous avons enregistré deux locutrices à la fois pour chaque session de jeu, chacune sur un canal séparé avec un micro-casque individuel. Afin d'assurer leur implication dans la tâche de jeu, nous avons sélectionné des femmes qui se considéraient comme compétitives et qui avaient l'habitude de jouer à des jeux avec leurs amis et/ou leur famille. Pour chaque binôme, les deux intervenants étaient des amis (un ou deux ans d'amitié en moyenne). Cette « condition d'amitié » visait à assurer le partage des connaissances, la complicité des joueuses et leur implication dans la tâche.

## 2.3 Matériel

204 mots bi-syllabiques du français comportant les voyelles /i, y, ɛ, œ, a, ɔ, o, u/ dans leur syllabe finale ont été sélectionnés pour constituer les mots cibles. Une contrainte supplémentaire était qu'ils devaient former des triplets liés sémantiquement (ex. : mot mystère, « livre », indice 1 « auteur », indice 2 « lecteur »).

Les contextes consonantiques des voyelles étant naturellement déséquilibrés en français, il n'a pas été possible de maintenir le contexte constant pour chaque catégorie vocalique lors du choix des mots bisyllabiques. Néanmoins, afin de limiter les variations non contrôlées liées à la coarticulation, nous avons essayé de réduire les différences autant que possible en privilégiant les consonnes coronales. Faute de place, les 22 contextes consonantiques ne seront pas listés ici. Les mêmes mots sont utilisés dans les quatre tâches, notre matériel est donc comparable dans toutes les conditions de production.

Bien que 91% des cibles aient été produites par les locutrices (la joueuse P2 devinait parfois avant le deuxième indice, ou l'indice était ignoré par P1), un seuil de 5 exemplaires par catégorie de voyelles par locutrices a été fixé, afin d'assurer la fiabilité de l'analyse. Nous avons donc exclu les voyelles mi-fermées /ø/ et /e/, ainsi que les productions d'une locutrice. Un total de 1702 exemplaires des voyelles /i, y, ɛ, œ, a, ɔ, o, u/ produites par 7 joueuses ont été analysés pour cette étude.

## 2.4 Analyses acoustiques

Les durées des voyelles et les valeurs de F1, F2 et F3 ont été extraites grâce à un script Praat, les valeurs pour chaque formant étant une moyenne de mesures prises au 1/3, 1/2 et 2/3 de la durée totale de la voyelle, suivant la méthode utilisé par Audibert et al. (2015). Les valeurs aberrantes ont été filtrées en suivant la procédure utilisée par (Gendrot et al., 2005), puis corrigées manuellement par examen des coupes spectrales. Après conversion en Bark, la variation des voyelles produites dans les différentes conditions a été exprimée en termes de :

(a) variabilité des réalisations acoustiques des exemplaires de voyelles dans chaque catégorie, interprétée comme une mesure de la précision et de la stabilité des cibles vocaliques. Pour chaque catégorie vocalique, une cible acoustique moyenne est calculée en termes de F1, F2 et F3. La variabilité intra-catégorie est mesurée par la distance euclidienne de chaque exemplaire par rapport au centroïde de sa catégorie.

(b) dispersion des exemplaires au sein du système vocalique dans l'espace F1*F2, mesurée comme la distance euclidienne de chaque voyelle au centroïde du système vocalique du locuteur. Cette mesure permet d'appréhender la dynamique de centralisation / expansion de l'espace acoustique

occupé par le système vocalique. (ici l'espace F1*F2*F3 ne nous a pas paru pertinent du fait du manque d'informations sur les possibilités d'interprétation d'une centralisation dans l'espace F2*F3).

# 3 Résultats

Les données ont été analysées avec un modèle linéaire mixte (Bates et al., 2015), dans lequel la variable dépendante était la métrique (a) ou (b), la catégorie vocalique (n= 8) et le style (les 4 conditions de production) des facteurs fixes, et le contexte consonantique et le joueur des facteurs aléatoires. L'inspection visuelle de la distribution des résidus ne révèle pas de violation claire des conditions de normalité et d'homoscédasticité. Des comparaisons par paires entre conditions de production ont été réalisées avec le test HSD de Tukey. La taille d'effet de chaque facteur fixe a également été estimée par les valeurs de $R^2$ marginal (Nakagawa et al., 2013).

Sans surprise, un effet significatif de la catégorie vocalique et du joueur ressort sur la durée des voyelles ($R^2$ = 8.7% et $R^2$ = 29%, respectivement), la dispersion des voyelles dans le système F1*F2 ($R^2$ = 24% et $R^2$ = 3.6%) et la variabilité intra-catégorie ($R^2$ = 11% et $R^2$ = 1.6%). Plus intéressant pour notre étude, le style agit aussi sur ces trois variables. La comparaison des tailles d'effet montre que l'effet du style est plus important sur la durée des voyelles ($R^2$=29%) que sur leur dispersion au sein du système ($R^2$=3.6%) ou la variabilité intra-catégorie ($R^2$=1.4%). Pour ces deux dernières dimensions, la figure 1 illustre la variation liée au style est comparée à la variation individuelle entre les joueurs. Il ressort que la dispersion au sein du système vocalique est relativement moins affectée par le style que par différences individuelles, tandis que la variabilité intra-catégorie est légèrement plus affectée par le style que par le joueur ou par l'interaction entre les deux variables.

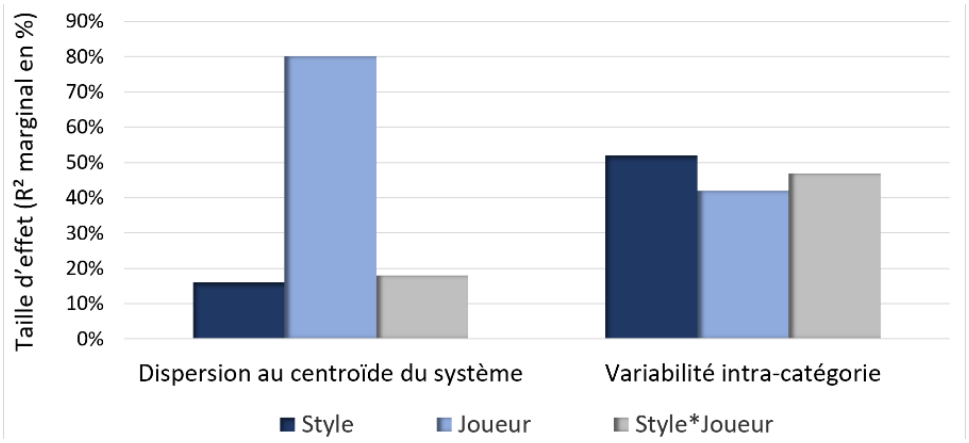



FIGURE 1: Comparaison de la taille de l'effet ($R^2$ marginal) des facteurs « style », « joueur », et de leur interaction, sur la dispersion au sein du système et les mesures de variabilité intra-catégorie.

La comparaison post-hoc entre les quatre styles de parole montre différents schémas selon la dimension étudiée. Les voyelles sont plus longues dans la condition de *clear speech* simulée (ME) que dans les trois autres conditions, comme illustré par la figure 2. Ce style de parole est également caractérisé par un espace acoustique F1*F2 plus grand avec des voyelles plus périphériques (i.e. moins centralisées) comme illustré sur la figure 3. La condition de lecture rapide (LR), au contraire, montre le schéma attendu pour le discours « hypo-articulé » comparé aux trois autres conditions : avec des voyelles plus courtes, et un espace acoustique F1*F2 plus petit avec des voyelles plus centralisées (cf. figures 2 et 3). Dans la condition de jeu (J) le patron est différent : les voyelles sont courtes comme en LR mais pas centralisées. En effet, l'espace acoustique dans la condition Jeu et

similaire à celui de la Lecture Normale (LN) comme illustré en figure 3. Concernant la variabilité intra-catégorie on observe deux tendances : la dispersion intra-catégorie est importante dans les conditions jeu et lecture rapide, alors qu'elle est faible dans les conditions ME et LN.

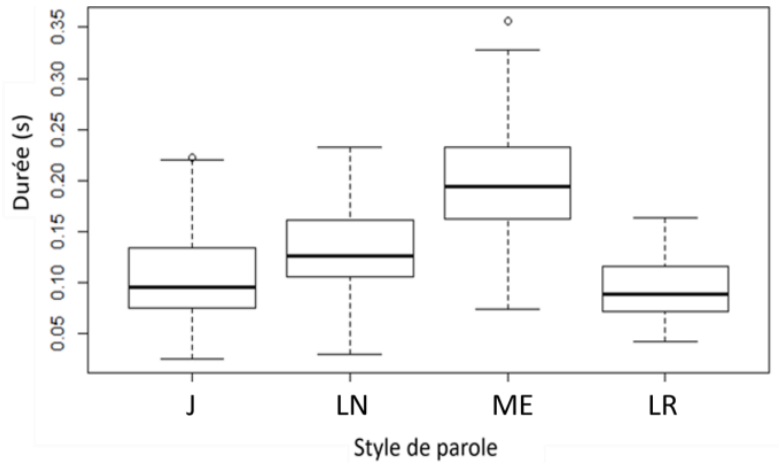



FIGURE 2: Durées vocaliques en fonction des conditions de la parole : Tâche de jeu (J) Lecture simple / normale (LN), lecture comme pour un auditeur malentendant (ME), et lecture rapide (LR).

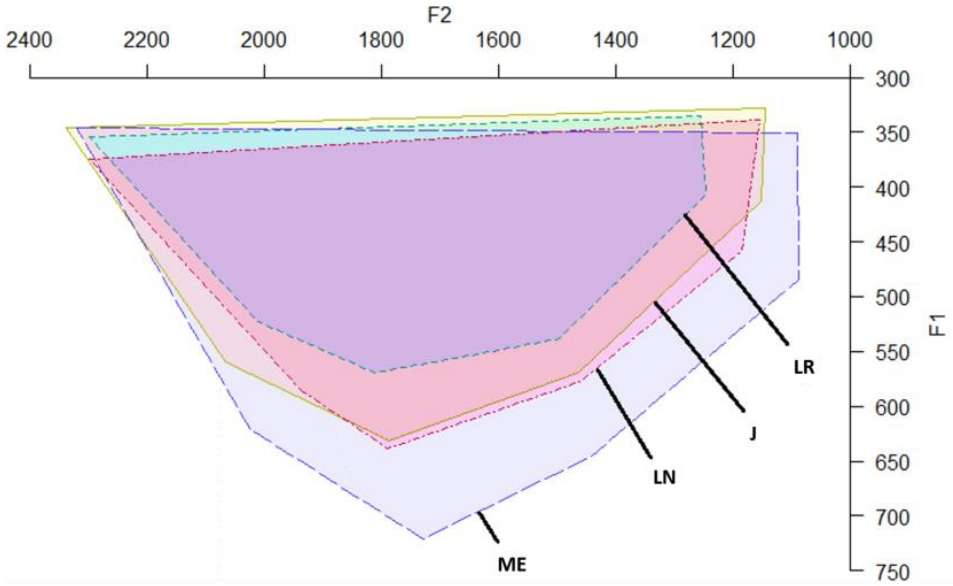



FIGURE 3: Représentation des espaces acoustiques de voyelle F1*F2 pour les styles jeu (J), lecture normale (LN), lecture à un malentendant (ME) et lecture rapide (LR). Le sommet des polygones représente la moyenne d'une catégorie vocalique toutes locutrices confondues.

Une analyse des corrélations entre la dispersion moyenne des exemplaires au sein du système et la variabilité intra-catégorie pour chaque locutrice dans les quatre condition de parole suggère une très faible interdépendance entre ces deux variables (avec au maximum r=0.2 en ME, et une forte variation inter-locuteurs).

La durée d'une voyelle a souvent été rapportée comme ayant une forte influence sur ses propriétés spectrales et notamment sur sa centralisation (Moon, 1989 ; Lindblom, 1991 ; Gendrot, 2005). C'est pourquoi nous avons tenu à tester également ce facteur. Nous avons donc calculé les corrélations entre l'indice de dispersion des exemplaires au sein du système et la durée, ainsi qu'entre l'indice de variabilité intra-catégorie et la durée dans nos quatre styles de parole. Les corrélations se sont révélées

très faibles pour tous les styles, indiquant également une faible interdépendance de ces deux facteurs dans nos données. Pour la dispersion par rapport au centre du système, la corrélation avec la durée est légèrement négative dans tous les styles, avec r = -0.2 en condition de lecture normale ($p<0.01$), r = -0.15 en lecture rapide ($p<0.01$), r = -0.25 pour la condition de lecture pour un mal entendant ($p<0.01$), et r = -0.1 pour le jeu ($p<0.01$). Les corrélations entre la variabilité intra-catégorie sont, elles de r= -0.22 pour LN ($p<0.01$), r= -0.14 pour LR ($p<0.01$), r= 0.08 en ME ($p>0.05$), et r= 0.07 en J ($p>0.05$).

# 4 Discussions et conclusions

En accord avec les travaux de la littérature (vus en introduction), nous constatons que la situation de production de la parole influence grandement la durée des voyelles. Dans une moindre mesure, des différences sont également observées pour la centralisation des voyelles dans l'espace F1*F2, et pour la variabilité entre les exemplaires d'une même voyelle (variabilité intra).

Comme prévu, les conditions ME et LR se situent aux deux extrêmes du continuum en termes de durée vocalique et de centralisation comme schématisé dans la Figure 5. Dans ces deux conditions, la variation spectrale liée à la durée peut être en jeu : des voyelles plus courtes et plus centralisées en condition LR *vs.* des voyelles plus longues et moins réduites en condition ME. Ce dernier résultat, réplique les variations observées par Scarborough et Zellou (2013) en anglais dans une même tache simulée 'comme à une personne malentendante' .

En ce qui concerne la parole interactive dans notre condition de jeu (J), nous observons un motif différent. Bien que les voyelles soient courtes, reflétant probablement un débit de parole plus rapide dans cette situation minutée, elles ne sont que peu réduites. Leur degré de centralisation n'est pas similaire à celui observé en condition LR, mais à celui observé pour la condition LN. Dans une étude antérieure basée sur de grands corpus incluant des locuteurs différents, Audibert et al. (2015), ont montré que la relation spatio-temporelle (ou compromis entre la vitesse et la précision acoustique) pouvait aussi différer entre « styles » de parole. Par exemple, les différences en termes de centralisation des voyelles entre discours « lu » et « journalistique » disparaissent lorsque la durée des voyelles est contrôlée et que les différences en termes de contrastivité des voyelles (estimées à partir des valeurs de F1 et F2) dépendent davantage de la durée des voyelles.

Les mesures de variabilité intra-catégorie montrent également l'intérêt d'examiner l'aspect multidimensionnel de la variation vocalique. La variabilité intra-catégorie est interprétée comme l'indexation de la précision des cibles vocaliques et de la variabilité de leur réalisation. Dans la condition de jeu, plus de variabilité intra-catégorie est trouvée et ce modèle n'est stable entre les locuteurs que dans cette condition de parole. De premières pistes d'explications peuvent être avancées. D'une part, des catégories plus précises et moins variables pourraient être un indice de « clarté » si ce phénomène était lié à des cibles plus contrastives montrant moins de chevauchement entre les catégories vocaliques adjacentes. Si tel était le cas, nous nous serions attendus à une réduction de la variabilité interne des catégories dans la condition ME, ce qui n'est pas le cas. D'autre part, l'augmentation de la variabilité entre les exemplaires de voyelles dans la situation de jeu pourrait être le signe d'une augmentation de la variation coarticulatoire. Or, la coarticulation est connue pour être dépendante du style et peut contribuer à augmenter l'intelligibilité, comme montré par Scarborough et Zellou (2013). Une autre direction à explorer serait que l'augmentation de la variabilité entre les exemplaires dans la condition de Jeu est due aux propriétés spécifiques de cette condition interactive, dans laquelle les affects et notamment les attitudes agissent également sur la réalisation acoustique des voyelles (voir par exemple Fónagy, 1983).

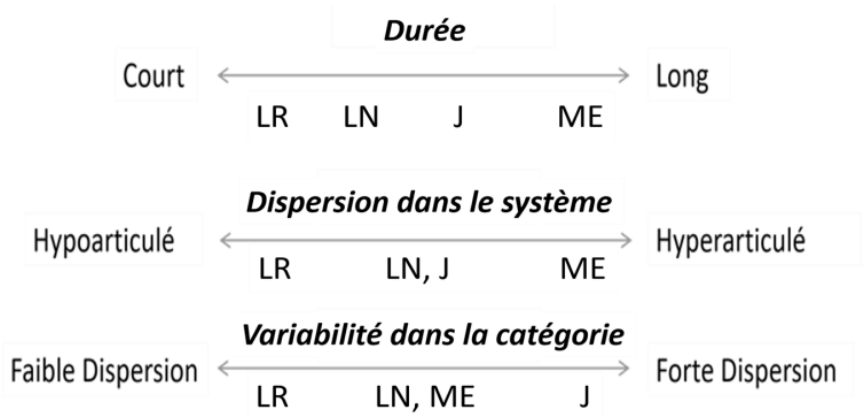



FIGURE 5: Répartition des styles de parole le long de chaque dimension acoustique.

# Remerciements

Cette étude a été soutenue par le Labex EFL (ANR-10-LABX-0083).

# Références

AUDIBERT N., FOUGERON C., GENDROT C., ADDA-DECKER M. (2015). Duration-vs. style-dependent vowel variation: A multiparametric investigation. Actes de *ICPhS 2015*.

BATES, D., MAECHLER, M., BOLKER, B., WALKER S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.

FERGUSON, S-H. et KEWLEY-PORT, D. (2007). Talker differences in clear and conversational speech: Acoustic characteristics of vowels. *Journal of Speech, Language, and Hearing Research*, 50 (5), 1241-1255.

FÓNAGY, I. (1983). *La vive voix : essais de psycho-phonétique* (Vol. 20). Payot.

FOUGERON C., et AUDIBERT N. (2011). Testing various metrics for the description of vowel distortion in dysarthria. Actes de *ICPhS* 2011, 687-690.

GENDROT C. et ADDA-DECKER M. (2005). Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German. *Variation*, 2(22.5), 2-4.

HARMEGNIES, B., & POCH-OLIVE, D. (1992). A study of style-induced vowel variability: Laboratory versus spontaneous speech in Spanish. *Speech communication*, *11*(4-5), 429-437.

HARMEGNIES B. et POCH-OLIVÉ D. (1994). Formants frequencies variability in French vowels under the effect of various speaking styles. *Le Journal de Physique IV*, *4*(C5), C5-509.

HUPIN B. et SIMON A-C. (2009). Analyse phonostylistique du discours radiophonique. Expériences sur la mise en fonction professionnelle du phonostyle et sur le lien entre mélodicité et proximité du discours radiophonique. *Recherches en communication*, 28, 103-121.

LINDBLOM B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech production and speech modelling* (pp. 403-439). Springer Netherlands.

LINDBLOM, B., BROWNLEE, S., DAVIS, B., & MOON, S. J. (1991). Speech Transforms. In *Phonetics and Phonology of Speaking Styles*.

MATHON C. (2014). Perception des phonostyles et représentativité du phonogenre : le cas du commentaire sportif en direct. *Nouveaux cahiers de linguistique française*, *31*, 93-103.

MOON, S. J., & LINDBLOM, B. (1989). Formant undershoot in clear and citation-form speech: A second progress report. *STL-QPSR*, *30*, 121-123.

NAKAMURA M., IWANO K., et FURUI S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language*, 22(2), 171-184.

NAKAGAWA, S., & SCHIELZETH, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in Ecology and Evolution, 4(2), 133-142.

OSGOOD C. E. (1962). Studies on the generality of affective meaning systems. *American Psychologist*, 17(1), 10.

ROUAS J-L., BEPPU M., et ADDA-DECKER M. (2010). Comparison of spectral properties of read, prepared and casual speech in French. In *Proceedings of LREC*.

SCARBOROUGH R. et ZELLOU G. (2013). Clarity in communication: "Clear" speech authenticity and lexical neighborhood density effects in speech production and perception. *Journal of the Acoustical Society of America*, 134(5), 3793-3807.

XXXIIe Journées d'Études sur la Parole
4-8 juin 2018, Aix-en-Provence, France

# Évaluation de l'adaptation par renforcement d'un générateur en langage naturel neuronal pour le dialogue homme-machine

Matthieu Riou Bassam Jabaian Stéphane Huet Fabrice Lefèvre
CERI-LIA, Université d'Avignon, Avignon, France
matthieu.riou@alumni.univ-avignon.fr, {bassam.jabaian, stephane.huet, fabrice.lefevre}@univ-avignon.fr

RÉSUMÉ

Jusqu'à récemment, la génération en langage naturel dans les systèmes de dialogue utilisait des systèmes à base de règles et de patrons, mais de nouveaux modèles à base de réseaux de neurones récurrents ont été proposés (Wen *et al.*, 2016a). Cependant ces modèles nécessitent une grande quantité de données d'apprentissage qui peuvent être compliquées à collecter et à annoter. Pour répondre à cette problématique, nous avons proposé un protocole d'apprentissage en ligne utilisant un algorithme de bandit contre un adversaire, permettant d'améliorer l'utilisation d'un modèle initial appris sur un corpus plus restreint généré par patrons (Riou *et al.*, 2017). Dans cet article, nous étudions l'intérêt pratique de notre approche en utilisant des données réelles obtenues par reconnaissance automatique de la parole des propositions des utilisateurs et en faisant évaluer les sorties du système par des humains.

ABSTRACT

**Evaluation of the reinforcement adaptation of a neural natural language generation system for human-machine dialogue.**

Traditional systems for natural language generation in spoken dialogue systems use patterns and rules to generate system answers. Recently, systems based on recurrent neural network models have been proposed (Wen *et al.*, 2016a). Those systems require a large amount of data to be learned, which can be difficult to collect and annotate. Therefore we proposed a framework to adapt the NLG module online through direct interactions with the users (Riou *et al.*, 2017). In this paper, we study the practical interest of the approach with real data collected as automatic speech recognition of users' suggestions and having humans assessing the system's outputs.

MOTS-CLÉS : génération en langage naturel, réseau de neurones récurrent, bandit contre un adversaire, apprentissage en ligne, adaptation à un utilisateur, reconnaissance automatique de la parole.

KEYWORDS: natural language generation, recurrent neural network, adversarial bandit, online learning, user adaptation, automatic speech recognition.

## 1 Introduction

Le composant de génération en langage naturel (NLG pour *Natural Language Generation*) d'un système de dialogue oral a pour rôle de transformer la réponse du gestionnaire de dialogue (qui est sous forme d'actes de dialogue) en une forme textuelle exploitable par le module de synthèse


10.21437/JEP.2018-40

vocale. Par exemple, l'acte $inform(type = restaurant, count = 4, food = pizza)$ peut générer les phrases « *There are 4 restaurants serving pizzas* » ou encore « *Pour manger de la pizza je peux vous indiquer 4 restaurants* ».

Les approches à base de patrons rédigés par des experts, traditionnellement utilisées pour cette tâche (Rambow *et al.*, 2001), produisent des phrases de bonne qualité pour des tâches bien spécifiées mais ces phrases restent assez répétitives et souvent peu naturelles. Les modèles statistiques permettent de palier ces limites avec des capacités de généralisation plus grandes, permettant l'introduction d'une plus grande variabilité lors de la génération mais aussi un plus grand risque d'erreur de syntaxe. Les premiers modèles de ce type utilisaient des modèles de langue n-grammes de mots (Oh & Rudnicky, 2002) et ont été étendus par la suite par l'utilisation de facteurs intégrant les étiquettes sémantiques dans le calcul des probabilités de séquences produites (Mairesse & Young, 2014). Plus récemment, avec l'expansion du *deep learning* dans différentes tâches de traitement du langage naturel, des modèles à base de réseaux de neurones récurrents ont été proposés pour le module de génération (Wen *et al.*, 2015a; Riou *et al.*, 2017). En parallèle, des nouvelles approches ont été étudiées pour apprendre des systèmes de bout en bout (Serban *et al.*, 2016). Ces études sont toutefois effectuées dans un contexte différent du notre car elles considèrent des systèmes conversationnels non dirigés par le but et pour lesquels de grandes ressources de données sont disponibles.

Afin de réduire le coût lié à une annotation d'un nouveau corpus de génération pour un nouveau domaine, Wen *et al.* ont proposé une approche incrémentale pour gérer l'adaptation de domaine d'un modèle de génération à base de RNN (Wen *et al.*, 2016b). Ils recourent à des données contrefaites synthétisées à partir d'un ensemble hors-domaine pour ajuster leur modèle sur un ensemble réduit de phrases du domaine. Il est aussi envisageable d'employer des méthodes d'extension de corpus (Manishina *et al.*, 2016) mais elles ne permettent pas une adaptation simultanée aux préférences de l'utilisateur.

Dans un travail précédent (Riou *et al.*, 2017), nous avons proposé de réduire le coût de la production de nouvelles données, en adaptant un modèle initial afin de générer des phrases avec une plus grande diversité. Dans cette logique, une approche par renforcement basée sur un algorithme de bandit contre un adversaire a été appliquée (Auer *et al.*, 2002) pour adapter un modèle RNN aux nouvelles phrases, différentes de l'ensemble d'apprentissage, en prenant en compte le coût imposé à l'utilisateur pour les fournir au système. Cette proposition a été évaluée sur des données simulant les erreurs de transcription automatique des retours des utilisateurs.

Dans cet article, nous étendons cette études en évaluant notre proposition sur des données réelles extraites d'un système de reconnaissance de la parole et en complétant les comparaisons préliminaires effectuées à l'aide de métriques automatiques par des mesures faites par des annotateurs humains sur trois critères : l'informativité, le naturel et la grammaticalité.

# 2 Le système de génération

Pour modèle de génération, nous avons proposé d'utiliser le LSTM (*Long short-term memory*) à contexte combiné décrit plus en détail dans (Riou *et al.*, 2017). Ce modèle combine deux modèles neuronaux proposés par Wen *et al.*, le modèle LSTM conditionné sémantiquement (SCLSTM) et l'encodeur-décodeur RNN avec attention. Chacun de ces systèmes traite différemment l'information sémantique représentée par un acte de dialogue (AD) pour produire la phrase. La cellule de contrôle (*reading gate*) du SCLSTM permet de filtrer l'AD en ne conservant à chaque étape que l'information restante, qui n'a pas été encore traitée. Au contraire, le mécanisme d'attention permet de sélectionner

dans l'AD au complet l'information à considérer spécifiquement à l'étape suivante, mais ne modélise pas explicitement la progression de l'information déjà traitée. L'objectif de notre système est de combiner les avantages des deux systèmes, en utilisant une cellule de contrôle similaire au SCLSTM et un mécanisme d'attention pour traiter séquentiellement l'AD restant à générer.

# 3 Un modèle pour l'adaptation en ligne

Afin de réduire le coût d'annotations complémentaires pour rajouter plus de variabilité dans le système, nous avons proposé un protocole d'apprentissage en ligne en deux étapes : tout d'abord, un modèle est appris sur un corpus constitué de références générées par patrons, puis le modèle est utilisé pour générer des phrases, en interaction vocale avec l'utilisateur, auquel il peut demander de produire un énoncé meilleur ou différent (Riou *et al.*, 2017).

On rappelle brièvement qu'à chaque tour de parole et après la génération de l'énoncé par le système, ce dernier doit décider de la meilleure action à suivre (à partir d'une distribution de probabilité et en tenant compte du coût $\phi(\text{Action})$ estimé) parmi :
• **Skip** : n'appliquer aucune mise à jour au modèle. Le coût de cette action est nul ($\phi(\text{Skip}) = 0$).
• **AskDictation** : affiner le modèle en considérant un énoncé alternatif proposé par l'utilisateur et transcrit automatiquement par un système de reconnaissance ($\phi(\text{AskDictation}) = 1$).
• **AskTranscription** : demander à l'utilisateur de transcrire la correction ou l'énoncé alternatif ($\phi(\text{AskTranscription}) = 1 + l$, avec $l$ la taille de l'énoncé proposé).

Nous avons estimé le **gain** de chaque action $g(i) \in [0, 1]$ comme décrit dans (Riou *et al.*, 2017) et nous avons défini une fonction de perte $l(i) \in [0, 1]$ qui permet de maximiser ce gain $g(i)$ tout en minimisant l'effort de l'utilisateur $\phi(i)$ :

$$l(i) = \underbrace{\alpha(1 - g(i))}_{\text{amélioration du système}} + \underbrace{(1 - \alpha)\frac{\phi(i)}{\phi_{max}}}_{\text{effort de l'utilisateur}} \quad (1)$$

avec $\alpha$ un scalaire qui pondère le gain par rapport au coût, permettant au système de s'adapter aux préférences de l'utilisateur et $\phi_{max}$ correspond à l'effort maximal pour normaliser l'effort de l'utilisateur entre 0 et 1.

Afin de réduire l'effort demandé à l'utilisateur et éviter des actions inutiles, un algorithme de bandit contre un adversaire à été adopté. A chaque itération, le système produit une phrase puis choisit une action $i_t \in \mathcal{I}$. Une fois que l'action $i_t$ est effectuée, le système calcule l'estimation du gain $g(i_t)$, l'effort de l'utilisateur $\phi(i_t)$ et la perte $l(i_t)$. Le rôle du bandit est donc de trouver $i_1, i_2, \ldots$, afin que pour chaque $t$, le système minimise la perte $l(i_t)$.

# 4 Expériences

## 4.1 Cadre expérimental

Les expériences ont été menées sur le corpus *SF restaurant* décrit dans (Wen *et al.*, 2015b) et librement accessible en ligne [1]. Ces données contiennent 5 191 phrases, pour 271 AD distincts. Le

1. https://www.repository.cam.ac.uk/handle/1810/251304

corpus associe à chaque acte une phrase générée par patron et plusieurs phrases proposées par des annotateurs humains, chaque phrase étant délexicalisée [2].

Le LSTM à contexte combiné a été implémenté en utilisant la bibliothèque Tensorflow [3]. Ce système a ensuite été entraîné sur le corpus séparé en 3 parties suivant un ratio 3 :1 :1 : apprentissage, validation et test, en utilisant uniquement les références proposées par les annotateurs humains.

## 4.2 Comparaison des systèmes de génération

Une évaluation a été conduite avec deux métriques objectives calculées à partir des générations de référence, le score BLEU-4 (Papineni *et al.*, 2002) et le taux d'erreur en concepts SER. Le BLEU-4 valide la génération de phrases, notamment la grammaticalité, tandis que le SER se concentre spécifiquement sur le contenu sémantique. Pour chaque exemple, nous produisons 20 hypothèses et ne gardons pour l'évaluation que les 5 meilleures selon le score donné par le modèle NLG. Des références multiples pour chaque AD sont obtenues en groupant les phrases délexicalisées du même AD et en les relexicalisant ensuite.

Le LSTM à contexte combiné obtient un score BLEU-4 de 71,1%, voisin des deux autres systèmes initiaux (72,2% pour le SCLTSM et 69,7% pour l'encodeur-décodeur), mais le taux d'erreur SER est divisé par trois par rapport aux autres systèmes, en passant de 0,77% et 0,65% pour le SCLSTM et l'encodeur-décodeur, à 0,24% pour le LSTM à contexte combiné. Cela veut dire qu'il propose une meilleure couverture des concepts à exprimer et donc moins d'omissions ou d'erreurs de concepts, ce qui est le principal but recherché pour un module NLG.

## 4.3 Évaluation de l'apprentissage en ligne

Nous avons utilisé le même corpus, mais avec un découpage en apprentissage, validation et test suivant un ratio 2 :1 :1. Le modèle NLG utilisé est encore le LSTM à contexte combiné. Pour initialiser le modèle, nous l'entraînons en utilisant les références générées par patron du corpus d'apprentissage. Le corpus de validation a permis de décider l'arrêt de la phase d'apprentissage (*early stopping*). Ensuite, nous avons simulé un apprentissage en ligne en réutilisant le corpus d'apprentissage, mais cette fois en apprenant sur les références proposées par des annotateurs humains. Le modèle ainsi que la politique de bandit ont été mis à jour toutes les 400 phrases. Dans cette série d'expériences, le WER a été simulé en insérant de manière aléatoire des erreurs (substitution, insertion et suppression) dans les exemples du corpus, jusqu'à atteindre un taux global de WER prédéfini.

Le modèle initial, entraîné sur les références générées par patron, atteint un haut score BLEU-4 de 80,2% quand celui-ci est calculé à partir de références générées par patrons, mais qui est réduit à seulement 39,7% en refaisant les calculs à partir des seules références proposées par des annotateurs humains. Cela montre la grande diversité possible des réponses dans une conversation.

## 4.4 Évaluation humaine

Les évaluations basées sur les métriques automatiques, tel BLEU-4, ne reflètent pas nécessairement les vraies préférences utilisateurs (Callison-Burch *et al.*, 2006). En particulier la dimension naturelle est très complexe à formaliser. Dans le but de mieux comparer les modèles initiaux et adaptés, nous

2. La délexicalisation remplace les formes de surface des concepts par des variables, *inform(name=la mimosa, food=mediterranean)* devient « *$name sert des plats $food* ».

3. `https://www.tensorflow.org`

| | Système initial | Système adapté |
|---|---|---|
| Score global | 2.356 | **2.425** |
| Informativité | **2.528** | 2.509 |
| Grammaticalité | 2.272 | **2.383** |
| Naturel | 2.267 | **2.383** |

TABLE 1 – Moyenne des scores pour chaque système

recourons à une évaluation humaine. 5 annotateurs ont reçu pour consigne d'évaluer les phrases générées automatiquement. Pour chaque exemple, l'évaluateur était confronté avec l'acte de dialogue visé et les 3 meilleures propositions de chaque système. Les hypothèses à juger sont ordonnées aléatoirement, les phrases équivalentes regroupées et aucune indication du système d'origine n'est disponible. Nous avons demandé aux annotateurs de donner à chacune des phrases (6 au maximum) trois scores :

- **Informativité** évalue si l'ensemble des informations présentes dans l'acte de dialogue sont bien toutes transmises dans la phrase générée, et si aucune supplémentaire n'est introduite, sur une échelle de 1 à 3 :
  - 3 : toutes les informations données par l'acte de dialogue (et seulement ces informations) sont présentes.
  - 2 : une information mineure est manquante, où une extra information non contradictoire est présente.
  - 1 : dans les autre cas.
- **Grammaticalité** évalue le niveau de correction syntaxique de la phrase, sur une échelle de 1 à 3 :
  - 3 : la phrase est correcte.
  - 2 : il y a quelques imperfections, mais qui probablement ne sont pas audibles.
  - 1 : il y a des erreurs importantes dans la phrase.
- **Naturel** évalue à quel point la phrase est proche d'une production potentielle humaine, sur une échelle de 1 à 3 :
  - 3 : la phrase aurait pu être prononcée par un humain dans cette situation.
  - 2 : la phrase est correcte mais moins appropriée à la situation, ou semble « automatique ».
  - 1 : même en corrigeant les erreurs grammaticales le cas échéant, la phrase n'aurait jamais pu être prononcée par un humain.

En supplément à l'évaluation de chaque propositions, les annotateurs ont aussi indiqué la phrase qu'ils jugeaient la meilleure de façon globale. Afin de mesure le niveau d'accord entre annotateur avec la métrique Kappa de Fleiss, les 20 premiers exemples étaient communs à tous. Au total, 471 annotations ont été réalisées

L'accord moyen global entre annotateurs présente un $\kappa$ de 0,55. La tâche qui présente le moins grand agrément est le jugement sur le naturel ($\kappa$=0.468), à comparer avec des $\kappa$ de 0.59 et 0.58 respectivement pour l'informativité et la grammaticalité.

Comme on peut le constater dans le tableau 1 le système adapté obtient un score global moyen plus élevé que le système initial. Plus particulièrement, ses scores sont meilleurs pour le naturel et la grammaticalité mais un peu dégradés pour l'informativité. Dans le tableau suivant 2, on peut observer que les deux systèmes ont tendance à avoir des scores globaux plus élevés pour des actes de dialogues de longueurs moyennes (2 à 3 slots). Au delà le score décroit rapidement du fait d'une plus grande

| # slots | Système | Tous | Informativité | Naturel | Grammaticalité |
|---|---|---|---|---|---|
| 0 | Initial | 1,74 | 1,76 | 1,72 | 1,74 |
| | Adapté | 2,59 | 2,60 | 2,60 | 2,58 |
| 1 | Initial | 2,38 | 2,59 | 2,34 | 2,21 |
| | Adapté | 2,38 | 2,55 | 2,31 | 2,27 |
| 2 | Initial | 2,58 | 2,80 | 2,50 | 2,46 |
| | Adapté | 2,64 | 2,75 | 2,58 | 2,58 |
| 3 | Initial | 2,47 | 2,72 | 2,30 | 2,39 |
| | Adapté | 2,32 | 2,48 | 2,26 | 2,29 |
| 4 | Initial | 2,28 | 2,35 | 2,24 | 2,27 |
| | Adapté | 1,71 | 1,74 | 1,69 | 1,70 |
| 5 | Initial | 1,75 | 2,00 | 1,67 | 1,58 |
| | Adapté | 1,66 | 1,50 | 1,75 | 1,58 |

TABLE 2 – Moyennes des scores pour chaque système en fonction du nombre de slots dans l'acte de dialogue

complexité, plus favorable à l'introduction d'erreurs dans le cas de la génération stochastique.

Enfin, le tableau 3 nous permet de constater les variations de scores en fonction des types d'actes de dialogue. On observe, contrairement à notre intuition, que les scores sont assez réguliers selon les types, et ce alors qu'ils représentent bien sûr des complexités très variables (mais qui doivent aussi être liées au nombre moyen de concepts associés en moyenne à chacun des actes, il peut par exemple être assez grand pour l'acte très générique inform, alors qu'il est nul pour goodbye).

Quand les annotateurs devaient voter pour leur phrase favorite, ils ont en majorité voté pour la meilleure proposition de chacun des systèmes (on rappelle que les phrases ne sont pas présentées de façon ordonnée), avec une préférence pour le système adapté (voir le tableau 4). Mais surtout on constate que les phrases proposées en positions 2 et 3 sont aussi beaucoup sélectionnées par les annotateurs dans le cas du système adapté, ce qui participe à confirmer que le système adapté peut générer des phrases satisfaisantes avec une plus grande variabilité que le système initial.

## 4.5 Évaluation de l'adaptation en ligne avec des vraies données orales

Pour évaluer en pratique le schéma proposé d'adaptation en ligne, et en particulier l'impact du taux d'erreur en mot de la reconnaissance de parole durant les interactions, une collecte de corpus a été réalisée.

Pour chaque phrase nous avons confronté un utilisateur avec son acte de dialogue (et la possibilité de lire quelques exemples de référence de génération de cette phrase, système initial). Puis l'utilisateur devait dicter une alternative correspondant à l'acte de dialogue, qui était automatiquement transcrite. Afin de simplifier le déploiement de l'expérience, les capacités de reconnaissance de la parole offerte par le navigateur Chrome, utilisant l'API RAP de Google, ont été utilisées. Enfin à partir de la sortie automatique l'utilisateur avait la possibilité d'apporter manuellement les corrections nécessaires pour fournir une transcription de référence. Les deux sorties, automatiques et références, ont été collectées, ainsi que le score de confiance des transcriptions automatiques.

| Acte de dialogue | Système | Tous | Informativité | Naturel | Grammaticalité |
|---|---|---|---|---|---|
| inform | Initial | 2.50 | 2,39 | 2,39 | 2,42 |
| | Adapté | 2,42 | 2.52 | 2,37 | 2,39 |
| inform_only_match | Initial | 2,33 | 2,50 | 2,50 | 2,00 |
| | Adapté | 1,94 | 2,00 | 2,00 | 1.83 |
| ?inform_no_match | Initial | 2,48 | 2,78 | 2,33 | 2,34 |
| | Adapté | 2,19 | 2,23 | 2,07 | 2,05 |
| ?select | Initial | 2,16 | 2,52 | 2,01 | 1,89 |
| | Adapté | 2.05 | 2,52 | 2,37 | 2,33 |
| ?request | Initial | 2,65 | 2,82 | 2,65 | 2.47 |
| | Adapté | 2,63 | 2,77 | 2,57 | 2,55 |
| ?reqmore | Initial | 2,27 | 2,67 | 2,07 | 2,07 |
| | Adapté | 2,62 | 2.47 | 2,73 | 2,67 |
| ?confirm | Initial | 2,02 | 2,27 | 1.91 | 1,88 |
| | Adapté | 2,05 | 2,33 | 1,94 | 1,88 |
| goodbye | Initial | 1,71 | 1,70 | 1,70 | 1,72 |
| | Adapté | 2,59 | 2,60 | 2,59 | 2,57 |

TABLE 3 – Moyennes des scores pour chaque système selon le type des actes de dialogue

| Rang | Système initial | Système adapté |
|---|---|---|
| 1 | 111 (22,0%) | 143 (28,5%) |
| 2 | 51 (10,2%) | 103 (20,6%) |
| 3 | 18 ( 3,6%) | 75 (15,0%) |
| Total | 180 (35,9%) | 321 (64,1%) |

TABLE 4 – Effectif de phrases sélectionnées par les annotateurs selon leur rang.

426 paires de transcriptions (automatiques, manuelles) ont pu être récupérées de cette manière. [4] Le taux d'erreur en mot moyen est de seulement 2,42%, avec un score de confiance moyen de 0,86.

Un nouveau modèle a été appris en suivant le protocole d'adaptation en ligne décrit dans la partie 4.3 (avec $\alpha = 0,5$) en utilisant les données nouvellement collectée, au lieu des références annotées par des humains. Le nouveau corpus est divisé en 300 phrases pour l'apprentissage et 126 pour le test. Nous avons gardé le modèle initial utilisé dans la première expérience, mais il a été cette fois mis à jour, ainsi que le bandit, toutes les 50 phrases du fait de la taille plus réduite du corpus. Pour améliorer l'estimation du gain, l'estimation du WER global a été remplacée par le score de confiance de la transcription :

$$g(\text{AskDictation}) = (1 - \text{BLEU}_{gen/prop}) \times \text{score de confiance} \times (1 - \text{SER})$$

Nous avons testé ce nouveau modèle sur le même corpus que dans la première expérience, en comparant aux références générées par patrons dans un premier temps, et dans un second temps aux références proposées par des annotateurs humains auxquelles nous avons rajouté nos propres références corrigées issues de la collecte de données orales. Les résultats montrent des tendances similaires aux résultats obtenus avec le WER simulé. Le score BLEU-4 par rapport aux références

4. Toutes les données utilisées dans cette étude sont disponibles sur demande.

générées par patrons diminue fortement (10%, de 82,9% à 72,7%), tandis qu'il chute légèrement lorsqu'on le compare aux références humaines (3%, de 48,17% à 45,08%), ce qui s'explique par la forte présence dans le test des références proposées par des annotateurs humains du corpus initial qui n'ont pas été apprises dans cette expérience.

L'algorithme de bandit permet de ne pas demander constamment des efforts importants à l'utilisateur. Sur la totalité de l'apprentissage il a demandé 53% du temps une transcription, contre 23% pour une alternative à l'oral et 23% aucune alternative. Cela lui permet de diviser presque par deux le coût cumulé sur l'apprentissage sans pour autant trop diminuer les performances par rapport aux références annotées par des humains, soit un coût cumulé de 2430 et un BLEU-4 final de 44,4% contre respectivement 4243 et 45,8% dans le cas où le système demanderait systématiquement des transcriptions. En revanche il fait moins bien que le système qui ne demanderait que des alternatives orales, qui obtiendrait un BLEU-4 équivalent de 45,1% pour un coût cumulé de 300.

Pour permettre au système de mieux évaluer s'il doit ou non risquer de demander une alternative orale, il faudrait pouvoir élargir le contexte de l'exemple traité pris en compte par le bandit (nature de l'acte de dialogue, complexité...), par exemple en gardant le même protocole mais en utilisant un algorithme de bandit contextuel (Auer *et al.*, 2002).

# 5 Conclusions et perspectives

Dans cet article, nous avons évalué un nouveau protocole pour adapter un modèle initial de génération de langage naturel neuronal à l'aide d'un apprentissage en ligne. Les résultats obtenus par une expérience simulée ont ainsi pu être confirmés et complétés avec des utilisateurs réels, pour fournir des propositions au système et juger des qualités des hypothèses du système adapté. L'algorithme de bandit permet d'équilibrer de manière automatique l'évolution des performances du système avec le coût induit pour l'utilisateur et d'aboutir à un système que les utilisateurs jugent plus varié. Une voie d'amélioration possible du système que nous entrevoyons est l'augmentation des capacités d'apprentissage du bandit par la prise en compte du contexte lors de ses décisions, à l'aide d'un bandit contextuel. Enfin, le générateur de texte doit pouvoir être évalué dans le contexte du système de dialogue complet pour confirmer l'intérêt pratique de l'approche.

# Remerciements

Ce travail a été partiellement financé par le Labex BLRI (ANR-11-LABX-0036) et l'ILCB.

# Références

AUER P., CESA-BIANCHI N., FREUND Y. & SCHAPIRE R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, **32**(1), 48–77.

CALLISON-BURCH C., OSBORNE M. & KOEHN P. (2006). Re-evaluating the role of bleu in machine translation research. In *EACL*, p. 249–256.

MAIRESSE F. & YOUNG S. (2014). Stochastic language generation in dialogue using factored language models. *Computational Linguistics*, **40**(4), 763–799.

MANISHINA E., JABAIAN B., HUET S. & LEFÈVRE F. (2016). Automatic corpus extension for data-driven natural language generation. In *LREC*.

OH A. H. & RUDNICKY A. I. (2002). Stochastic natural language generation for spoken dialog systems. *Computer Speech & Language*, **16**(3–4), 387–407.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *ACL*.

RAMBOW O., BANGALORE S. & WALKER M. (2001). Natural language generation in dialog systems. In *HLT*.

RIOU M., JABAIAN B., HUET S. & LEFÈVRE F. (2017). Online adaptation of an attention-based neural network for natural language generation. In *INTERSPEECH*.

SERBAN I. V., SORDONI A., BENGIO Y., COURVILLE A. & PINEAU J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI Conference on Artificial Intelligence*.

WEN T.-H., GAŠIĆ M., MRKŠIĆ N., BARAHONA L. M. R., SU P.-H., ULTES S., VANDYKE D. & YOUNG S. (2016a). Conditional generation and snapshot learning in neural dialogue systems. In *EMNLP*.

WEN T.-H., GAŠIĆ M., KIM D., MRKŠIĆ N., SU P.-H., VANDYKE D. & YOUNG S. (2015a). Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. In *SIGDIAL*.

WEN T.-H., GAŠIĆ M., MRKŠIĆ N., ROJAS-BARAHONA L. M., SU P.-H., VANDYKE D. & YOUNG S. (2016b). Multi-domain neural network language generation for spoken dialogue systems. In *NAACL-HLT*.

WEN T.-H., GAŠIĆ M., MRKŠIĆ N., SU P.-H., VANDYKE D. & YOUNG S. (2015b). Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *EMNLP*.

# L'histoire des alphabets phonétiques du XVIIIe siècle jusqu'à l'API

Claudia Schweitzer[1], Christelle Dodane[2], Jan Lazar[3]
(1) Histoire des théories linguistiques (HTL), UMR 7597, Université Paris 3 Sorbonne nouvelle, CNRS, et Université Montpellier 3, Département musicologie, Université Paul-Valéry, Route de Mende, 34199 Montpellier cedex 5, France
(2) PRAXILING, UMR5267, Université Paul Valéry Montpellier 3, CNRS, Bâtiment Marc Bloch (BRED), Route de Mende, 34199 Montpellier cedex 5, France
(3) Université d'Ostrava, République Tchèque et Université d´Opole (Pologne)

claudia.schweitzer2@gmail.com, christelle.dodane@univ-montp3.fr, jan.lazar@osu.cz

## RESUME

Au cours de l'histoire, il y a eu de nombreuses tentatives de créer un alphabet permettant d'établir l'inventaire des sons des langues du monde et ce, dès le XVII[e] jusqu'à l'apparition à la fin du XIX[e] de l'Alphabet Phonétique International (API, 1888). Pour atteindre cet objectif, nous savons aujourd'hui que chaque symbole doit désigner un seul son et inversement. Mais ce principe ne s'est pas imposé d'emblée et il a fallu de nombreuses tentatives avant d'y arriver. C'est cette histoire que nous allons tenter de retracer en exposant les réflexions de Charles de Brosses (Alphabet organique, 1765) et de Montmignon (Alphabet figuré, 1785), puis les différentes tentatives de création d'alphabets au XIXe siècle qui conduiront à la création de l'API en 1888 (Alphabet phonotypique de Pitman et Ellis, 1845, Visible Speech de Bell, 1867, Palaeotype de Ellis, 1869 et Alphabet romique de Sweet, 1877).

## ABSTRACT

**The history of alphabets from the 18th century to the IPA**

In the course of history, there have been many attempts to create an alphabet to establish the inventory of sounds of the languages of the world, from the eighteenth until the appearance in the late nineteenth century, of the International Phonetic Alphabet (IPA, 1888). To achieve this goal, we now know that each symbol must designate a single sound and vice versa. But this principle did not prevail from the beginning and it took many attempts to get there. This is this story we will try to trace by exposing the reflections of Charles de Brosses (Organic Alphabet, 1765) and Montmignon (figurative Alphabet, 1785), and then, the various attempts of creation of alphabets during the nineteenth century that will lead to the API in 1888 (Ellis and Pitman's Phonotypic Alphabet, 1845, Bell's Visible Speech, 1867, Ellis's Palaeotype, 1869 and Sweet's Romanic Alphabet, 1877).

MOTS-CLES : Histoire, alphabets, phonétique, API.
KEYWORDS: History, alphabets, phonetics, IPA.


10.21437/JEP.2018-41

# 1 Introduction

L'histoire de la description de la langue ou, autrement dit de sa grammaticalisation, est étroitement liée à l'écriture. C'est grâce à elle que l'on peut établir des normes et les règles d'une langue. L'écriture développe donc toute sa puissance dans la transmission des informations et réflexions, c'est ce que les anciens auteurs appellent souvent les « idées » (voir par exemple la *GGR* 1660). Pourtant, dans notre culture occidentale, l'écriture alphabétique est censée représenter la langue orale en premier lieu. Pour ce faire, elle impose des signes et suppose l'existence d'un système : un alphabet, faisant la liste de tous les signes à la disposition du scripteur. Ainsi, le dictionnaire de Furetière (1690, article « alphabet ») va définir l'alphabet comme la *« disposition par ordre des lettres d'une langue ».* L'enjeu consiste alors à dresser la liste complète des unités nécessaires pour noter de manière cohérente les énoncés oraux. Mais l'entreprise n'est pas aussi facile qu'il n'y paraît. D'un côté, il s'agit certes de disposer d'un assez grand nombre de signes.[1] Mais d'un autre côté, un trop grand nombre de signes rend les deux actes de l'écriture et de la lecture compliqués. C'est le moment où la limitation, voire la réduction du nombre de signes entre en jeu. Au XVII[e] siècle encore, Ménage (1675 : 454), dans ses *Observations sur la langue françoise*[2] évoque la difficulté ainsi que l'importance de la réflexion sur ce fait : *« Je remarqueray icy en passant, que c'est bien plus d'ajouter une lettre l l'Alphabet, que d'ajouter un mot à la Langue ».* On en trouve encore les traces un siècle plus tard dans l'article « alphabet » de L'*Encyclopédie* où les auteurs exposent : *« Par alphabet d'une langue, on entend la table ou liste des caracteres, qui sont les signes des sons particuliers qui entrent dans la composition des mots de cette langue. Toutes les nations qui écrivent leur langue, ont un alphabet qui leur est propre, ou qu'elles ont adopté de quelque autre langue plus ancienne. Il seroit à souhaiter que chacun de ces alphabets eut été dressé par des personnes habiles, après un examen raisonnable ; il y auroit alors moins de contradictions choquantes entre la maniere d'écrire & la maniere de prononcer, & l'on apprendroit plus facilement à lire les langues étrangeres. »* (ENC, art. alphabet). Depuis les premières grammaires du français au XVI[e] siècle jusqu'aux Encyclopédistes, un problème se pose effectivement aux théoriciens de la langue : le français dispose de davantage de sons qu'il n'y a de caractères dans l'alphabet latin. Par conséquent, ni la prononciation d'un texte écrit, ni la transposition de la langue parlée en écriture ne sont évidentes, et ce fait est vrai autant pour les Français, que les apprenants étrangers. Un deuxième problème va occuper les phonéticiens du XIX[e] siècle : le fait qu'une écriture alphabétique ne note pas les sons physiques, mais qu'elle reproduit des phonèmes. En effet, selon Alan Kemp (2001) la plus grande difficulté dans la recherche d'un nouveau système de notation phonétique réside dans la recherche de différents symboles rendant possible l'identification des sons, de façon non ambigüe, étant donné qu'aucun alphabet existant ne le permette. Pour atteindre cet objectif, chaque symbole doit désigner un seul son et pour faciliter la lecture, les symboles doivent être facilement différenciés, tout en s'harmonisant les uns aux autres. Enfin, le système doit pouvoir être étendu pour intégrer de nouveaux sons et de nouvelles langues. C'est seulement en remplissant ces conditions qu'il pourra avoir une portée générale et universelle. Selon lui, les tentatives de création d'alphabets peuvent se classer en trois groupes différents : 1) les systèmes dits iconiques, morphologiques ou physiologiques, 2) les systèmes alphabétiques et 3) les systèmes non alphabétiques qui rejettent toutes formes existantes au profit de nouvelles formes. Au cours de cet article, nous nous intéresserons principalement aux deux premières catégories, en décrivant deux exemples d'alphabets qui ont été mis au point au cours du XVIII[e] siècle, l'alphabet figuré de Montmignon (1785) et l'alphabet organique de Charles de Brosses (1765), ainsi que les différents systèmes de notation qui ont mené à la création de l'Alphabet Phonétique International en 1888 : l'alphabet phonotypique de Pitman et d'Ellis (1845), le « Visible Speech » de Bell (1867), le

[1] Dans cet article, nous utilisons le terme « signe » pour la figure qui représente dans les alphabets matériellement le son (ou le phonème) de la langue parlée.

[2] Il s'agit de remarques à l'instar de celles publiées par Vaugelas en 1647.

Palaeotype d'Ellis (1869) et l'alphabet romique d'Henry Sweet (1877). Dans cet article, nous renvoyons à différentes figures que l'on peut consulter à partir du lien suivant : https://drive.google.com/open?id=16hRRYZoCmwpUVVNIBpbdANgmarigJAsc.

# 2 Les alphabets pour le français au XVIII[e] siècle

Après plus de deux siècles pendant lesquels les auteurs français avaient discuté sur les lettres de l'alphabet et sur leur réalisation notamment sur le fond de la non-correspondance entre l'écrit et l'oral, à partir de 1740 environ (moment où les philosophes intègrent l'Académie Française), les réflexions prennent une tournure plus générale. On continue à réfléchir sur les moyens de normer l'orthographe française d'une part, de l'autre on cherche à découvrir la généralité des sons en comparaison avec les autres langues connues. Ce travail est évidemment fortement influencé par la linguistique comparative naissant en Allemagne. Deux exemples peuvent illustrer ces efforts : l'alphabet figuré de Montmignon et l'alphabet organique de de Brosses.[3]

## 2.1 L'alphabet figuré de Montmignon (1785)

L'alphabet figuré de Jean-Baptiste Montmignon (1785) se comprend comme une contribution à la langue nationale (française). Il veut rendre la prononciation du français univoque, et cela pour plusieurs groupes de personnes : d'abord les Français mêmes (pour éviter la prononciation dialectale et pour faciliter l'apprentissage pour les enfants), puis pour les apprenants étrangers, et enfin aussi pour les Français vivant à l'étranger qui *« ont perdu l'accent national, & qui l'abâtardissent, en le mélant avec l'accent du pays qui les a adoptés »* (1785 : 75). L'auteur développe alors un alphabet qui veut rendre la prononciation des mots écrits univoque et cela d'abord pour le français, mais aussi pour d'autres langues. L'idée est de *« peindre les sons de la parole »* et de *« rapprocher le plus qu'il se pourra, la Langue écrite de la Langue parlée, de manière que les yeux lisent dans l'écriture, tout ce que la prononciation fait entendre à l'oreille »* (1785 : 86). Chez Montmignon, un son simple correspond à un signe simple, un son composé (comme une diphtongue) à un signe composé. L'auteur propose ensuite d'ajouter aux caractères représentant les sons vocaliques (simples et composés) des signes diacritiques supplémentaires donnant des informations sur le timbre (ouvert ou fermé), la longueur (longue, brève ou moyenne) ou la nasalité des sons. Pour écrire avec cet alphabet, il faut d'abord diviser le mot en syllabes. Cette opération sert à savoir plus tard quels sont les sons à rassembler ensemble lors de la lecture. Sous chaque son vocalique prononcé, on place le signe indiquant sa réalisation exacte (par exemple un *e* ouvert, long et oral sera noté par une ligne descendante, un *e* ouvert, bref et oral, par une ligne montante et un *e* ouvert, moyen et oral, par une ligne verticale). Les consonnes non prononcées seront marquées par des signes de retranchement (-) sous les caractères concernés. Les signes consonantiques correspondant à plusieurs prononciations possibles, sont dotés d'une marque pour indiquer leur prononciation réelle (ainsi, dans le mot *oser*, la prononciation du *s* en [z] doit être indiquée). Dans un tableau ajouté à la fin du texte, Montmignon adapte également les nouveaux signes à la langue anglaise (*cf.* Figure 1).

## 2.2 L'alphabet organique de Charles de Brosses (1765)

L'alphabet organique de Charles de Brosses (1765) repose sur la comparaison des langues et l'observation de l'acquisition du langage par les enfants. De Brosses développe une théorie

[3] On peut voir ce développement dans le contexte des travaux sur la constitution d'une grammaire générale, applicable à toutes les langues. Ils débutent en 1660, se développent au XVIII[e] siècle et en France, deviennent très importants au XIX[e] siècle.

phonomimétique (Droixhe 1978 : 192), selon laquelle l'ordre des sons appris par l'enfant reproduit en quelques mois le grand cycle de l'érudition du langage humain. Il obéit à des règles articulatoires privilégiant les sons faciles à former aux sons demandant une grande flexibilité articulatoire et les sons simples aux combinaisons de plusieurs consonnes. L'enfant commence alors par les sons bilabiaux, suivis des occlusives vélaires, des apico-dentales, des alvéo-dentales et des palatale. Dans un premier tableau (*cf.* Figure 2) établi par l'auteur, on voit que le signe stylisé essaie de rendre (grossièrement) compte de ces traits. De Brosses essaie ensuite de développer un modèle plus abstrait par la réduction des images en signes géométriques. L'alphabet développé sur cette base repose sur l'analyse des mouvements articulatoires : une ligne droite indique les labiales, une ligne descendante les palatales et une ligne montante les gutturales. Les labiales, les palatales et les nasales sont indiqués par une hampe au bout supérieur de la ligne qui, quant à elle, peut également être droite ou penchée. Les signes se distinguent ensuite par des petits traits diagonaux ou des petits points pour différencier les modes et les lieux d'articulation. Enfin, toutes les voyelles sont indiquées par une ligne verticale, dotée de petites lignes horizontales à différents endroits, indiquant par sa hauteur le lieu d'articulation du son. La formation des sons est considérée comme un processus mécanique, fait dont les signes ou *« lettres organiques »* (Séris 1995 : 280) rendent compte. Dans ce sens, l'alphabet de Charles de Brosses est universel : il peint la production des sons, et il propose alors des signes dont le résultat (complètement mécanique) peut être supposé être sans équivoque.[4]

# 3 Vers la création d'un alphabet universel au XIXe siècle

Le XIX$^{e}$ siècle est traversé par la volonté de créer un système de transcription phonétique unique pour toutes les langues. Cette volonté transparaît dans la création du prix Volney en 1822, dont l'objectif est d'atteindre la création d'un ensemble de principes destinés à réduire les langues orales à une écriture alphabétique en caractères latins. La tentative la plus célèbre est celle de l'égyptologue, philologue et archéologue Karl Richard Lepsius (1855) pour l'écriture des langues du monde et en particulier les langues d'Afrique, basé sur l'alphabet latin (appelé alphabet standard ou alphabet général de Lepsius). Il a tenté de décrire les différences essentielles entre les sons, plus d'une cinquantaine selon lui, mais au final, son système de notation comprend au moins 286 caractères. Cette volonté de créer un alphabet phonétique rejoint la volonté de créer des langues universelles permettant à l'ensemble de la population mondiale de pouvoir communiquer. La fin du XIX$^{e}$ siècle voit ainsi la création de plusieurs langues artificielles à visée universelle (le solresol par François Sudre en 1866, l'esperanto en 1887 par Ludwig Zamenhof et le Volapük en 1879-1880 par Johan Martin Schleyer). Johan Martin Schleyer avait notamment souligné la nécessité d'un alphabet phonétique international permettant de retranscrire les sons de toutes les langues du monde. Les tentatives les plus fécondes furent élaborées au cours de la seconde moitié du siècle lorsque la connaissance des sons devient plus approfondie et plus scientifique.

## 3.1 L'alphabet phonotypique d'Isaac Pitman et Alexander John Ellis (1845)

L'alphabet phonotypique représente une tentative de simplification de l'orthographe de l'anglais grâce un système de notation phonétique et un alphabet qui permet de représenter chaque son par une seule lettre. Il a été utilisé dans l'un des premiers dictionnaires phonétiques de l'anglais. Il utilise les mêmes phonèmes que la méthode Pitman Shorthand, un système sténographique créé par Issac Pitman en 1837 et largement diffusé aux Etats-Unis et en Grande-Bretagne. L'alphabet phonotypique représente une extension de l'alphabet latin qui inclue des signes de ponctuation. Il a été développé

[4] En outre, l'alphabet organique est un instrument étymologique qui permet de *« mesurer le degré de comparaison entre les langues »* (de Brosses 1765 : 177) par le nombre des parentés des sons qui forment les mots. L'alphabet organique du président de Brosses s'intègre de cette façon dans le courant de la grammaire comparée.

par Isaac Pitman et Alexander John Ellis en Angleterre en 1844 et la première version stable a été publiée en 1847. Le terme phonotypique a été utilisé pour différencier ce système du terme phonographique utilisé par Pitman pour le shortland[5]. Il a été par la suite appelé alphabet phonétique anglais, mais son alphabet n'est pas phonétique dans le sens employé aujourd'hui. Certains de ses éléments seront incorporés à l'API. La première version contenait 40 lettres, permettant de distinguer les voyelles courtes des voyelles longues (*cf*. Figure 3), mais non les syllabes accentuées des syllabes non accentuées. Quelques lettres additionnelles et beaucoup de combinaisons de signes diacritiques ont été conçues pour permettre l'écriture d'autres langues telles que l'allemand, l'arabe, l'espagnol, le florentin, le gallois, l'italien, le néerlandais, le polonais, le portugais ou encore le sanskrit.

## 3.2 Le Visible Speech d'Alexander Melville Bell (1867)

En 1867, le linguiste écossais Alexander Melville Bell (1867), père d'Alexander Graham Bell, l'inventeur du téléphone, va créer un système iconique composé de symboles qui indiquent la position et le mouvement de la gorge, de la langue et des lèvres lorsqu'elles produisent les sons du langage. Il a créé ce système pour aider les sourds à parler et notamment son propre fils. Comme chez de Brosses, il s'agit d'un système basé sur la physiologie car chaque signe porte en lui des informations visuelles sur la façon de le prononcer. Bell utilise dix symboles radicaux (*cf*. Figure 4) à partir desquels tous les autres symboles sont formés. Ainsi, la 4$^{e}$ consonne en partant de la gauche sur la seconde ligne est formée par les symboles radicaux 6, 8 et 1 (*cf*. Figure 5). Son système est modulaire dans le sens où les symboles utilisés pour les voyelles se différencient visuellement des symboles utilisés par les consonnes (*cf*. Figure 5). A cette base, Bell a ajouté un ou plusieurs éléments qui ne concernent qu'une toute petite partie de la lettre. La connaissance des éléments composant les différents symboles permet au lecteur d'interpréter l'ensemble des symboles (Bell, 1867 : 838). Ainsi, les symboles iconiques utilisés pour les consonnes peuvent contenir les informations suivantes : la lettre renversée C indique qu'il s'agit d'une consonne, la partie potentiellement ouverte représente les points correspondant sur le palais avec indication du lieu d'articulation (alignement entre le symbole consonantique et la ligne supérieure, à droite) et de la vibration des plis vocaux à l'aide d'une petite barre verticale (*cf.* Figure 5). Le problème, c'est que ces différents éléments sont très difficiles à identifier et qu'il y a donc un problème de lisibilité, notamment au niveau des voyelles. Il semble donc qu'il ne faille pas faire reposer des différences majeures de sons sur des petits détails typographiques. Chaque son doit avoir son propre symbole, ce qui évite de multiplier le nombre total de symboles. Bell a d'ailleurs parlé d'alphabet phonétique général à propos de son système de notation, car il peut en effet être utilisé pour noter n'importe quelle langue et il permet de transcrire beaucoup plus de nuances de prononciation que l'API (Mac Mahon, 1996 : 838). Si le système de Bell a eu beaucoup de succès à l'époque aux États-Unis, son manque de lisibilité et la difficulté de l'enseigner sans passer par un enseignement oral (Sweet, 1877 : 10) en ont limité la portée. L'un de ses élèves, Henry Sweet a tenté d'améliorer le système de notation de Bell en créant l'alphabet organique révisé (1880-1881), avec des symboles plus facilement lisibles. En Angleterre, ce système amélioré a été bien vite plus utilisé que celui de Bell[6]. Selon Mac Mahon (1996 : 171), le système de notation Visible Speech a donné lieu à l'analyse en traits distinctifs, les sons de la parole étant analysés dans leurs différentes composantes articulatoires.

[5] La sténographie shortland est un système créé par Pitman en 1837 pour permettre la transcription rapide de l'anglais oral. Il s'agit du système de sténographie le plus utilisé aujourd'hui au Royaume-Unis et aux Etats-Unis. Il est fondé sur un système phonétique où les symboles ne représentent pas des lettres, mais des sons.

[6] Un autre alphabet organique a été mis au point par Paul Passy et Daniel Jones en 1907, avec une ressemblance partielle avec ceux de Bell et de Sweet (Mac Mahon, 1996 : 840).

## 3.3 Le Palaeotype dialectal d'Alexander John Ellis (1869)

Il s'agit d'un alphabet phonétique mis au point par Alexander John Ellis pour décrire les sons des différents dialectes de l'anglais dans les années 1880. « Palaetotype dialectal » signifie vieil alphabet. Il se distingue des alphabets précédents, notamment de l'alphabet phonotypique créé par le même auteur par l'usage de l'alphabet latin et le fait que les différents caractères peuvent tous se trouver dans les casses des imprimeries (*cf.* Figure 6). Il utilise notamment des caractères en italique, des petites capitales et lettres inversées. Les formes (h, j, *w*) sont utilisées comme signes diacritiques, alors que les formes en capitales (H, J, w et q) représentent les sons consonantiques. Les voyelles longues sont représentées en doublant leur signe et les diphtongues par leur voyelle élémentaire en succession immédiate. Si elles ne forment pas de diphtongue, elles sont séparées par une virgule. Il contient un répertoire de plus de 250 caractères désignant des sons différents. Sa principale avancée est l'adaptation de l'alphabet latin ordinaire pour la représentation de différentes nuances de sons, sans avoir recours à des signes diacritiques comme dans le cas de l'alphabet général de Lepsius, très difficile à utiliser selon Sweet (1877, preface, vii-viii). Ellis précise que son palaeotype a surtout été réalisé à des fins scientifiques, pour décrire les sons entendus en anglais et dans les autres langues. Il donne la correspondance entre les symboles utilisés dans son système de notation et celui de Bell (Critical Notices, The North American Review, 1870 : 421). Cet alphabet va inspirer l'alphabet romique d'Henry Sweet qui va servir de base à l'API.

## 3.4 L'alphabet romique d'Henri Sweet (1877)

Henry Sweet est un fervent admirateur du système de notation mis au point par son ancien professeur Alexander Melville Bell, le « Visible Speech ». Non seulement, il va proposer une révision de cet alphabet organique (1880-1881), mais avant cela, il va mettre au point un nouvel alphabet, dans lequel il va retenir la terminologie de Bell qu'il estime remarquablement claire et concise (Sweet, 1877, preface, x, xi). Cet alphabet appelé romique est fondé, comme le pateotype, sur l'utilisation des caractères de l'alphabet latin. Il expose cette nouvelle notation en détail dans le chapitre IV « *Sound Notation* » de son livre « *A Handbook of Phonetics* » (1877, 100-168). Selon Sweet, sans un système clair de notation, il est impossible de discuter des questions de phonétique de façon pertinente ou de décrire la structure phonétique du langage (1877 : 100). Selon lui, il existe plusieurs solutions pour contourner les imperfections des caractères latins : utiliser 1) de nouveaux types de caractères, 2) des diacritiques comme l'accent grave ou aigu, 3) des digraphes (th, kh, etc.), 4) des lettres retournées, des italiques et des capitales. Les deux premières solutions sont critiquables car il faut d'abord utiliser les moyens existants avant d'en créer de nouveaux. Il va ainsi écarter entièrement le phonotype de Pitman et l'alphabet standard de Lepsius, les deux systèmes les plus connus relevant des deux premiers principes (1, 2) et avoir recours aux 2 principes suivants (3, 4) qui emploient les caractères ordinaires utilisés dans les imprimeries, comme dans le Paelotype d'Ellis (1877 : 101). Selon lui, il est nécessaire d'avoir un alphabet qui indique seulement les distinctions fondamentales qui correspondent aux distinctions de sens dans les langues et les noter par des symboles qui peuvent être facilement écrites et mémorisées. Sweet cherche à concevoir un système général qu'il est possible de modifier selon des principes définis pour s'adapter aux langues spécifiques. Dans son ouvrage, il donne ainsi des exemples d'application de l'alphabet romique à l'anglais, au français, à l'allemand, au hollandais, à l'icelandais, au vieil icelandais, au suédois et au danois. Comme l'avait fait Ellis, il ajoute des digraphes et des lettres retournées, ainsi que des caractères en italique. Il emprunte des lettres à l'anglo-saxon (e dans l'a : ae, eth) ou du grec (thêta) et renonce à l'utilisation des lettres majuscules. Par ailleurs, il instaure une distinction claire entre deux types de transcription phonétique : la transcription « *romique large* » qui permet de représenter les sons des différentes langues avec les mêmes symboles (*cf*. Figure 7) et la transcription « romique étroite » (*cf*. Figure 8) qui permet de représenter les spécificités phonétiques d'une langue en particulier et dont les caractères

sont notés en italique. Selon Anderson (1985 : 172), Sweet n'utilise pas encore le terme de phonème, mais il différencie ces deux types de transcrition sur un principe phonémique, le premier servant à différencier les sons sur une base phonémique alors que le second, sert à l'analyse fine des sons dont les propriétés phonétiques ne servent pas à distinguer le sens. Par ailleurs, Sweet abandonne la division en mots et la remplace par une division en groupes accentuels où les accents sont indiqués par des symboles différents. Sweet utilise également des symboles pour transcrire les variations de hauteur et de la qualité de la voix.

# 4 La naissance de l'alphabet phonétique international (1888)

L'alphabet romique de Henry Sweet va poser les bases de l'API. En effet, Sweet va relater son expérience de la transcription large avec l'alphabet romique aux membres de l'Association des Professeurs d'Anglais, une association qui réunissait au départ les professeurs d'anglais de six pays différents. Celle-ci avait été fondée deux années auparavant par Paul Passy en 1886 et commença à publier en mai de la même année un mensuel « *The Phonetic Teacher* » entièrement rédigé en transcription phonétique (*cf*. Figure 9). Elle va progressivement s'élargir aux professeurs d'autres langues (Galazzi, 1994). En 1889, elle prit le nom d'« *Association Phonétique des Professeurs de Langues Vivantes* » et en 1897, selon les vœux d'Otto Jespersen, d'« *Association Phonétique Internationale* ». Selon eux, il est nécessaire d'adopter un alphabet unique pour toutes les langues et universel, dans le sens où il puisse représenter toutes les articulations possibles des sons des langues humaines et d'en faciliter l'enseignement et l'apprentissage (Durand, à paraître). Le système de Sweet sera adopté sans grandes modifications par les membres de l'Association en juillet 1888, alors qu'auparavant, chacun utilisait des systèmes différents. On parlera désormais d'Alphabet Phonétique International. Ils ont l'idée de s'inspirer de la phonologie et vont choisir les caractères selon un principe phonémique, c'est-à-dire une lettre pour un phonème. Cela suppose l'étude préalable des différentes langues et la connaissance des phonèmes propres à chaque langue. Par ailleurs, pour que l'alphabet soit applicable à toutes les langues, il faut le généraliser. Comme les organes de la parole sont les mêmes pour la production des sons des différentes langues, on peut s'appuyer sur des critères articulatoires et notamment le mode d'articulation, le lieu d'articulation et le mode de phonation. Ces trois critères sont encore aujourd'hui à la base du classement de la charte de l'API (dernière version actualisée en 2015). Selon Durand (à paraître), d'autres critères de base ont été conservés jusqu'à aujourd'hui outre le principe phonémique et la généralisation : le principe de similitude selon lequel deux sons identiques dans deux langues différentes seront désignées par le même symbole, le principe d'universalité selon lequel c'est l'usage le plus courant dans les langues qui dictera l'adoption d'un nouveau symbole en utilisant l'alphabet latin, le principe d'iconicité selon lequel l'adoption de nouvelles lettres doit suggérer les sons par similitude aux lettres déjà existantes et le principe unitarien selon lequel les signes diacritiques doivent être évités dans la mesure du possible. Comme c'était déjà le cas dans le Visible Speech de Bell, l'API repose sur une forte distinction entre les symboles utilisés pour les consonnes et les voyelles. Comme l'alphabet figuré de Montmignon, l'alphabet phonotypique de Pitman & Ellis, le système palaeotype d'Ellis et l'alphabet romique de Sweet, l'API repose sur les caractères de l'alphabet latin et l'ajout de petites capitales, de lettres de type cursive, de nombreux diacritiques, de lettres retournées et de digraphes. Aujourd'hui, l'API est l'alphabet phonétique et phonologique le plus répandu dans le monde. Il permet de décrire la plupart des langues du monde et même de compléter certaines orthographes manquantes, comme dans le cas de l'alphabet pan-nigérian. Cependant, à cause de l'utilisation des caractères latins, le principe d'universalité est à remettre en question car la transcription de langues comme le français, l'anglais, l'espagnol et toutes les langues occidentales en général est plus aisée que celle d'autres langues, comme les langues asiatiques par exemple du fait de l'utilisation de caractères de l'alphabet latin. Selon Allard (2010 : 93), il ne s'agit donc pas réellement d'un principe universel, mais occidental car la moitié de la

population mondiale n'a pas d'écriture à base latine. Pour ces populations, il est donc nécessaire d'apprendre dans un premier temps l'alphabet latin, puis d'apprendre à l'adapter pour avoir accès à l'API.

# 5 Discussion et conclusion

Nous avons essayé de montrer avec l'exemple du français la problématique à laquelle se voient généralement exposées les langues disposant d'une écriture : la non-correspondance, plus ou moins importante, entre les signes alphabétiques dont se compose l'écriture et leur prononciation. En France, les longs débats sur l'orthographe – qui d'ailleurs sont encore loin d'être terminés – ont poussé certains auteurs à réfléchir à une transcription plus adéquate, plus facile, et notamment plus universelle des sons à l'aide d'alphabets. La volonté de généraliser la langue française à d'autres langues y joue évidement un rôle primaire. Mais sur le plan international, les recherches vont se focaliser au XIX$^{e}$ siècle sur la recherche d'un alphabet phonétique universel. On peut ainsi comparer les approches phonotypiques de Montmignon, de Pitman & Ellis et d'Ellis qui souhaitent améliorer la correspondance entre l'oral et l'écrit. Leurs travaux présentent de nouvelles versions de l'alphabet latin, reposant sur l'utilisation de lettres connues, comme ce sera notamment le cas plus tard avec les voyelles de l'API, elles aussi dotées d'indications précises pour leur prononciation exacte. Si l'alphabet de Montmignon ayant recours aux signes diacritiques n'est pas véritablement praticable car il suppose pour chaque mot deux transcriptions (une « normale » et l'autre figurée), dans ceux de Lepsius et d'Ellis, le nombre trop élevé de signes (plus de 250 caractères) s'avère également un obstacle à leur mise en pratique. La solution de Charles de Brosses et de Bell repose sur la théorie phonomimétique et représente la volonté de donner des indications sur la formation des sons. Comme l'alphabet est organique, voire mécanique, elle est censée être applicable à toutes les langues. Il s'avère pourtant que la praticabilité de ce type d'alphabet s'avère restreinte. Non seulement par le fait qu'il est nécessaire d'apprendre un système d'écriture entièrement nouveau, mais de plus, l'identification rapide des caractères n'est pas facilitée car les signes utilisés sont trop proches et leurs distinctions trop fines. La révision de Sweet de l'alphabet de Bell dans le but d'améliorer sa lisibilité passe alors de nouveau par le recourt à un alphabet basé sur des caractères latins, mais avec des digraphes et des lettres retournées, des italiques et des capitales pour indiquer les détails concrets de leur prononciation et de leur accentuation. Son alphabet romique se distingue par sa précision, indiquant de manière facilement mémorisable les distinctions fondamentales entre les sons et leur notation en lettres. Si tous les systèmes décrits vont apporter leur contribution à la création de l'API, ils apportent également très concrètement des réflexions et des expériences. L'avantage des systèmes utilisant les lettres latines (connues des lecteurs et écrivains, disponibles de surplus dans les cases des imprimeurs) est évident. Par ailleurs, il est facile de les étendre pour évoluer pour intégrer de nouveaux sons. En revanche, on peut remettre en question la portée universelle de tels alphabets étant donné qu'ils restent difficiles à lire et à appliquer dans les autres langues non occidentales. Les signes diacritiques, utilisés déjà par Montmignon pour caractériser les sons, ont finalement été adoptés dans l'API, par exemple pour l'indication des nasales ou la notation de la quantité. Comme c'était déjà le cas dans le Visible Speech de Bell, l'API repose sur une forte distinction entre les symboles utilisés pour les consonnes et les voyelles. Comme les alphabets phonotypique, palaeotype et romique, l'API utilise les caractères de l'alphabet latin auxquels il ajoute divers petits signes de distinction. Le gain par le retour des alphabets organiques est pourtant également clair : ils ont donné lieu à une analyse de plus en plus fine des traits distinctifs, c'est-à-dire des composantes articulatoires des sons. L'API se situe donc dans la continuité et dans l'héritage des travaux des grammairiens et phonéticiens.

# Remarques

Cet article dans une version plus approfondie et couvrant une période plus large incluant la Renaissance fait partie d'un projet de livre, en cours de réalisation, intitulé « *Histoire de la description de la parole : de l'introspection à l'instrumentation* » organisé par Christelle Dodane et Claudia Schweitzer et dont la publication est prévue d'ici la fin de l'année 2018.

# Références


ALLARD, G. (2010). *Composantes graphiques des systèmes phonétiques et leurs influences sur l'apprentissage et la compréhension du langage*. Mémoire de 5ème année, École Supérieure d'Art et de Design d'Amiens.

ANDERSON, S. (1985). *Phonologie in the Twentieth Century*. Chicago: The University of Chicago Press.

AUROUX, S. (1994). *La révolution technologique de la grammatisation*. Liège : Mardaga.

AUROUX, S., DESCHAMPS, J., KOULOUGHI, D. (2004). *La philosophie du langage*. Paris : PUF.

AUROUX, S. (1994). *Histoires des idées linguistiques*. Sprimont : Mardaga (Tome 3).

BELL, A. M. (1867). *Visible Speech: The science of Universal alphabetics*. London: Simpkin, Marshall & Co.

CRITICAL NOTICES (1870). *Alexander J. Ellis Early English Pronunciation*. The North American Review, Volume 0110, Issue 227 (April 1870), Art. VII, 420-438.

DE BROSSES, C. (1765). *Traité de la formation méchanique des langues et des principes physiques de l'étymologie*. Paris : Saillant-Vincent-Desaint (vol. 1).

DROIXHE, D. (1978). *La Linguistique et l'appel de l'histoire (1600-1800). Rationalisme et révolutions positivistes*. Paris : Droz.

DUMARSAIS, C. C., MALLET, E.-F. (1751). Article « Alphabet », L'Encyclopédie », 1e édition, Diderot, d'Alembert, vol. 1, 295-297.

DURAND, J. (A PARAITRE). L'alphabet phonétique international. Dans Herrenschmidt, C., Mugnaioni, M.J., Savelli, M.J., Touratier, C. (éds.). *Le Monde des Écritures*. Paris : Gallimard.

FURETIERE, A. (1690). *Dictionnaire universel*. La Haye : Leers.

GALAZZI, E. (1994). L'association phonétique internationale. Dans Auroux (éds.). *Histoire des idées linguistiques*. Sprimont : Mardaga, 499-516, (Tome 3).

KEMP, A. (2001). The history and development of a universal phonetic alphabet in the 19th century: from the beginnings to the establishment of the IPA. *An International Handbook on the Evolution of the Study of Language from the Beginnings to the Present. History of the Language Sciences*. New York: De Gruyter, Vol. 2, 1572-1595.

MAC MAHON, M. (1996). Phonetic notation. Dans Daniels P.T., Bright, W. *The World's writing systems*. New-York: Oxford University Press, 821-846.

MENAGE, G. (1675). *Observations de Monsieur Ménage sur la langue française.* Paris : C. Barbin.

MONTMIGNON, J.-B. (1785). *Système de prononciation figurée applicable à toutes les langues*. Paris : Royez.

QUEMADA, B. (1968). *Les Dictionnaires du français moderne 1539-1863*. Paris : Didier.

SERIS, J.P. (1995). *Langages et machines à l'âge classique*. Paris : Hachette.

SWEET, H. (1877). *A handbook of phonetics, including a popular exposition of the principles of spelling reform*. Oxford: Clarendon Press.

# Une histoire des JEP : 50 ans d'études sur la parole

Véronique Delvaux[1] Giancarlo Luxardo[2] Fabrice Hirsch[2]

(1) FNRS & IRSTL, UMONS, Belgique
(2) Laboratoire Praxiling, UMR5267 CNRS, Université Paul-Valéry Montpellier 3, France
veronique.delvaux@umons.ac.be, giancarlo.luxardo@univ-montp3.fr,
fabrice.hirsch@univ-montp3.fr

## RESUME

Cet article retrace l'histoire des *Journées d'Etudes sur la Parole* en mettant en avant l'évolution de cette manifestation scientifique au cours de ses 50 ans d'existence, et ce à différents niveaux: (i) l'organisation (lieux, durée, format des communications et des actes, structuration en différentes sessions, etc.); (ii) les participants et auteurs des communications (nombre, sexe, champ disciplinaire); (iii) les thématiques et méthodologies déployées dans les articles complets, sur la base d'une analyse textuelle de l'ensemble des actes publiés (31 éditions, près de 2000 articles complets).

## ABSTRACT

**The JEP: 50 years of research on speech**
This paper recounts the history of the *Journées d'Etudes sur la Parole,* outlining the evolution of this scientific event over its 50 years of existence. Several changes are scrutinized: (i) the organization (places, duration, format of proceedings and communications, session arrangement); (ii) the participants and authors of communications (number, gender, disciplinary field); (iii) the topics and methodologies in the full papers, based on a content analysis of all the published proceedings (31 editions, around 2000 full papers).

MOTS-CLES : JEP, histoire, sciences de la parole, textométrie
KEYWORDS : JEP, history, speech sciences, textometry.

# 1 Introduction

L'étude de l'histoire des revues ou des colloques ayant marqué de leur empreinte une discipline scientifique est une thématique en pleine émergence (par ex. Guérin-Pace *et al.*, 2012 pour un historique d'une revue de géographie ; Sturm, 2015). Dans le domaine de la parole, Sturm (2015) a retracé l'évolution du Congrès International des Sciences Phonétiques (International Congress of Phonetic Sciences). Cette étude a notamment permis de relever une évolution des présentations proposées lors de cette manifestation scientifique d'un point de vue méthodologique : les recherches réalisées sur la parole sont en effet progressivement devenues plus quantitatives au fil des années et utilisent désormais davantage les statistiques pour valider leurs résultats.


10.21437/JEP.2018-42

L'histoire des Journées d'Etudes sur la Parole (JEP) a également donné lieu à un certain nombre d'études (Boë & Liénard, 1988 ; Grossetti, 1994 par ex.). D'après Grossetti (1994), les prémices des JEP se situent à la fin des années 1960. C'est en 1967 en effet que les institutions grenobloises s'intéressant à la parole, accompagnées d'audiologistes lyonnais, organisent un colloque sur "les structures acoustiques de la parole". Cette manifestation, qui se déroule à Grenoble, regroupe plus d'une centaine de participants autour de 23 communications portant sur des thématiques liées à la phonétique, la linguistique, la psychologie, la physiologie, l'électronique, l'informatique, les télécommunications et l'acoustique. Suite à ce colloque, qui est le premier du genre à réunir des spécialistes des sciences de la parole, auront lieu un an plus tard, en 1968, les premières journées informelles d'étude sur la parole, organiséees par M. Wajskop, déjà présent l'année précédente à Grenoble, en collaboration avec le Département de Phonétique de Londres ainsi que des représentants des laboratoires de Paris, d'Aix-en-Provence et de Grenoble. Selon Carré (1973), ces premières journées sont déjà placées sous le signe de la pluridisciplinarité en vue d'étudier la parole. En 1970, ont lieu les premières "Journées d'études sur la parole" officielles. Celles-ci se tiennent à Grenoble, sous la responsabilité de J.P. Tubach (CETA), R. Carré (ENSERG-LCP), M. Wajskop (directeur de l'Institut de Phonétique de Bruxelles), M. Rossi (Institut de Phonétique d'Aix-en-Provence), P. Simon (Institut de Phonétique de Strasbourg) et R. Lancia (ENSERG). Contrairement aux manifestations précédentes, les JEP de 1970 font la part belle cette fois à l'informatique et à l'électronique avec des présentations portant notamment sur la synthèse et la reconnaissance vocales. Dans la foulée, se crée, à l'intérieur du "Groupement des Acousticiens de Langue Française" (GALF), le groupe "Communication Parlée" (GCP) réunissant les chercheurs travaillant dans le domaine de la parole. Dès lors, les JEP sont étroitement associées à leur société savante de référence (GCP puis GFCP et enfin AFCP[1]).

A propos de la première rencontre, qui a eu lieu à Bruxelles en 1968, René Carré précise: "Nous avions été accueillis royalement. Les Journées étaient très chargées, à cause du travail, et les nuits très courtes, pour d'autres raisons. Je vous conseille par exemple, une dégustation de gueuze ou bien une soupe à l'oignon vers 4h du matin, ou bien un bon repas de moules." (1973, p.11). La belle alliance entre travail scientifique et atmosphère festive était lancée... Cela étant, si l'esprit de convivialité qui anime ces rencontres a traversé le temps, il n'en demeure pas moins que les Journées d'Etudes sur la Parole ont évolué depuis leurs origines, témoignant ainsi des changements qu'a connu la communauté *Parole* ces cinquante dernières années.

Dès lors, l'objectif de cet article est de retracer l'évolution des JEP au cours du temps, en examinant leurs aspects logistiques et organisationnels, en s'intéressant aux auteurs de communication(s), en

[1] Les mutations ne manquent pas : en 1986, le GALF devient la SFA ("Société Française d'Acoustique), et en 1988 naît l'ESCA ("European Speech Communication Association"), ancêtre de l'ISCA, en partie sous l'impulsion du GCP. Le GFCP ("F" pour "Francophone") est dès lors, dès 1990 un groupe spécialisé à la fois de la SFA et de l'ESCA. En 2002, le GFCP s'affranchit de la SFA pour devenir une société savante à part entière, l'AFCP ("Association Francophone de la Communication Parlée"), elle-même un "Special Interest Group" de l'ISCA ("International Speech Communication Association").

particulier à quelques "grands noms" qui ont marqué de leur empreinte l'histoire de cette manifestation, et en étudiant les thématiques abordées dans les travaux publiés à la suite des Journées. Pour mener à bien cette étude, l'intégralité des Actes des JEP a été téléchargée sur le site de l'Association Francophone de la Communication Parlée (http://www.afcp-parole.org/spip.php?rubrique27). Les textes ont ensuite été rassemblés dans un corpus permettant notamment d'en extraire des données quantitatives par date ou par auteur, mais aussi de réaliser une étude textométrique afin de décrire l'évolution des sujets traités.

# 2 Organisation des Journées d'Etudes sur la Parole[2]

## 2.1 Localisations et (co-)organisations

L'organisation de chaque édition a lieu dans une ville différente, et est confiée à une équipe d'organisateurs locaux, disposant eux-mêmes de relais au sein de l'association. Au départ, les JEP ont lieu tous les ans et sont organisées par un seul laboratoire. A partir de la fin des années 1980, le rythme devient bisannuel, et les JEP sont régulièrement co-organisées par plusieurs équipes ou laboratoires, parfois issus de villes ou d'universités différentes (par ex. 1987: ENIT, IRSIT& IBLV, Tunis et LIMSI, Paris). En 2016 à Paris, les JEP ont été organisées par des équipes et chercheurs provenant de 15 laboratoires d'Ile-de-France !

Dès le début, les JEP dépassent les limites de l'hexagone et occupent le territoire de la francophonie : elles sont organisées en 1973, 1984 et 1992 à Bruxelles ; en 1981 et 1990 à Montréal ; en 1987 à Hammamet (Tunisie) ; en 1998 à Martigny (Suisse) ; enfin plus récemment en 2004 à Fès (Maroc) et en 2010 à Mons (Belgique). L'année 2004 inaugure par ailleurs la collaboration avec l'ATALA (Association pour le Traitement Automatique des Langues) : une édition sur deux, soit tous les 4 ans, les conférences JEP et TALN sont organisées conjointement, sur un même site, et partagent des activités communes (conférences invitées, événement sociaux, ateliers, etc.).

## 2.2 Format des Actes et des communications

Le format des Actes témoigne de l'évolution technologique des outils de la recherche ces 40 dernières années. Au départ, les Actes étaient édités en un ou deux volumes, sous forme imprimée uniquement, et les différents articles ne répondaient pas à un format standardisé. L'édition 2000 (organisée par l'ICP à Aussois) marque un tournant : chaque édition des JEP dispose désormais d'un site internet, où les Actes sont rendus disponibles sous forme électronique. Un CD-rom accompagne les Actes papier

[2] Pour rédiger cette section, on a eu recours non seulement aux actes eux-mêmes, mais aussi à l'information récoltée sur les sites internet ou dans la documentation diffusée lors des différentes éditions des JEP, voire dans les mémoires des organisateurs... Nos remerciements à tous les contributeurs.

jusqu'en 2010. En 2012, une clé USB contenant les articles complets accompagne un livret de résumés. Depuis 2014, les Actes des JEP sont totalement dématérialisés, soit disponibles uniquement sur internet. Le format des communications, et plus généralement la structuration de la conférence en diverses sessions, a connu beaucoup de variations au cours de ces cinquante ans. Au départ, les JEP accueillent uniquement des communications orales, en nombre limité, mais faisant l'objet de longues discussions, synthétisées et mises en perspectives par un rapporteur puis dûment rapportées dans les actes. Ces premières éditions sont souvent organisées autour de quelques thématiques fortes, sans viser l'exhaustivité[3].

Au fil des années, le nombre de communications augmente régulièrement. Avant 1980, on ne dépasse pas une trentaine d'articles par édition des JEP (alors que les participants à l'édition de 1972 sont déjà une centaine). Dans les années 1980, on atteint environ soixante-dix articles, puis une centaine dans les années 1990. Le record est atteint en 2006 (Dinard) : 124 des 164 soumissions sont acceptées pour présentation cette année-là. Le taux d'acceptation est difficile à établir avant 2000, il oscille entre 65% et 75% ces quinze dernières années. Les participants dits "payants" sont entre 100 et 150 les années "JEP" et entre 250 et 350 les années "JEP-TALN".

## 2.3 Organisation des sessions

Etant donné le nombre croissant de soumissions de qualité, le format de la conférence a évolué au fil du temps. L'organisation en sessions parallèles n'ayant pas convaincu (format testé à Bruxelles en 1984), les JEP se tournent vers un format plus long (elles durent 2 ou 3 jours avant 1985, 4 ou 5 jours après), où se côtoient communications orales et posters. La première session affichée aux JEP date de 1980 (Strasbourg). La répartition des communications est longtemps relativement équivalente : une moitié est présentée sous forme orale, une moitié sous forme affichée. Depuis une dizaine d'années, le rapport est plutôt d'un tiers (orales) - deux tiers (posters). Notons par ailleurs qu'on observe dans les années 2010 un fléchissement des soumissions - et donc des communications - aux JEP, témoin sans doute du nombre accru de conférences intéressant les chercheurs du domaine, ainsi que d'une réduction des ressources mises à leur disposition pour voyager.

Tout au long de leur histoire, les JEP sont caractérisées par une grande diversité dans le format des sessions : tables rondes, sessions thématiques avec rapporteur ou non, conférences invitées, conférences grand public, sessions spéciales, puis plus récemment, notamment suite à la collaboration avec l'ATALA, tutoriaux, sessions de démonstrations et ateliers. L'évolution des thématiques associées à ces sessions reflète les mutations du domaine : par ex. 1976 (Nancy): "Aide aux handicapés", "Transcription graphème-phonème", 1984 (Bruxelles): "L'information linguistique

[3] Par exemple, l'édition de 1977 à Aix-en-Provence, est centrée sur des thématiques relevant des sciences humaines (principalement, en prosodie) alors que l'édition de 1979 à Grenoble ne regroupe que des travaux de synthèse (et plus marginalement, de reconnaissance) de la parole, sans aucune session consacrée à la phonétique.

contenue dans le signal acoustique: analyse et invariance", 1996 (Avignon): "Ressources linguistiques", 2000 (Aussois): "Expertises vocales", 2016 (Paris): "Langue écrite, parlée, signée".

Ainsi, les modalités d'organisation des JEP ont évolué au fil des éditions. La partie suivante vise à documenter les participants et auteurs de communication en vue de savoir si leur profil a également évolué au cours du temps.

# 3 Participants et auteurs de communication

Nous étudions tout d'abord le profil des premiers auteurs de chaque article publié lors des éditions des JEP de 1974, 1984, 1994, 2002 et 2014, en termes de sexe, de laboratoire d'origine et de champ disciplinaire. Nous examinons ensuite les méta-données concernant l'ensemble des communications publiées (près de 2000 en 31 éditions), plus spécifiquement le nombre d'auteurs par article et le nombre moyen d'articles par auteur, avec un intérêt plus particulier pour les plus grands contributeurs.

## 3.1 Le sexe des participants

La Figure 1 porte sur le sexe des premiers auteurs de chaque article publié lors des éditions des JEP de 1974, 1984, 1994, 2002 et 2014. Ces Actes ont été sélectionnés dans le but d'observer l'évolution des JEP tous les 10 ans. Signalons cependant que nous avons mené cette étude avec l'édition des JEP de 2002, celle de 2004 étant co-organisée avec l'ATALA. Si les hommes étaient largement majoritaires durant les premières éditions (91% d'hommes *vs*. 9% de femmes en 1974), le ratio entre les deux sexes tend à s'équilibrer à mesure que l'on avance dans le temps, les Journées de 2014 comptant 53% d'hommes et 47% de femmes premier auteur.

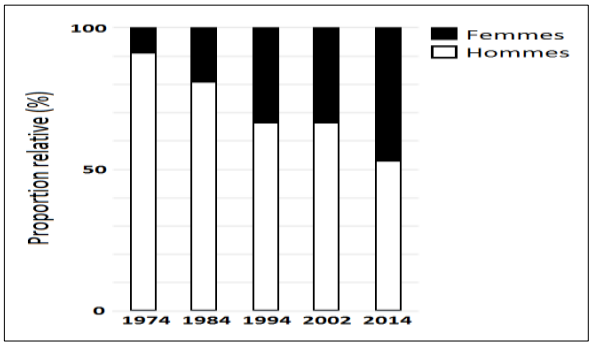



*Figure 1. Proportion relative de femmes et d'hommes dans les premiers auteurs d'articles publiés dans les Actes des JEP en 1974, 1984, 1994, 2002 et 2014.*

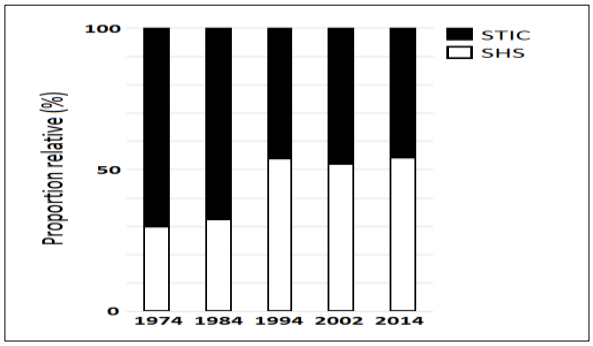



*Figure 2. Proportion relative de chercheurs en STIC et SHS dans les premiers auteurs d'articles publiés dans les Actes des JEP de 1974, 1984, 1994, 2002 et 2014.*

## 3.2 Participants STIC *vs*. SHS

La communauté *Parole* présentant la particularité d'être composée à la fois de chercheurs en SHS et en STIC, nous examinons la répartition de ces deux catégories d'intervenants à travers différentes éditions. La Figure 2 porte sur le champ disciplinaire du premier auteur des articles présents dans les 5 éditions des Actes sélectionnées. On constate que dans les premières années, les chercheurs étaient davantage issus des STIC par rapport aux SHS, les premiers cités représentant 70% des communications en 1974 *vs*. 30% pour les seconds. Ce rapport tend à s'équilibrer avec un léger avantage pour les SHS à partir des années 90, puisqu'en 1994 (de même qu'en 2014), 54% des premiers auteurs proviennent des sciences humaines et sociales et 46% des sciences et technologies de l'information et de la communication.

En résumé, sur base des éditions sélectionnées, on peut affirmer que le profil des chercheurs participant aux Journées d'Etudes sur la Parole a évolué en 50 années d'existence. Si les années 70 semblaient se caractériser par la présence d'une majorité d'hommes issus de laboratoires en STIC, les éditions futures ont vu l'arrivée de davantage de femmes et de chercheurs en SHS[4].

## 3.3 Nombre moyen d'auteurs par article

La Figure 3 dévoile le nombre moyen d'auteurs par article pour chacune des 31 éditions des Actes depuis la création des JEP. Les articles sont majoritairement écrits à moins de 2 contributeurs de 1970 à 1984. A partir de cette date, les papiers sont rédigés en moyenne par deux auteurs ou plus. A partir de 2004, les articles sont régulièrement rédigés à trois contributeurs ou plus (Figure 3).

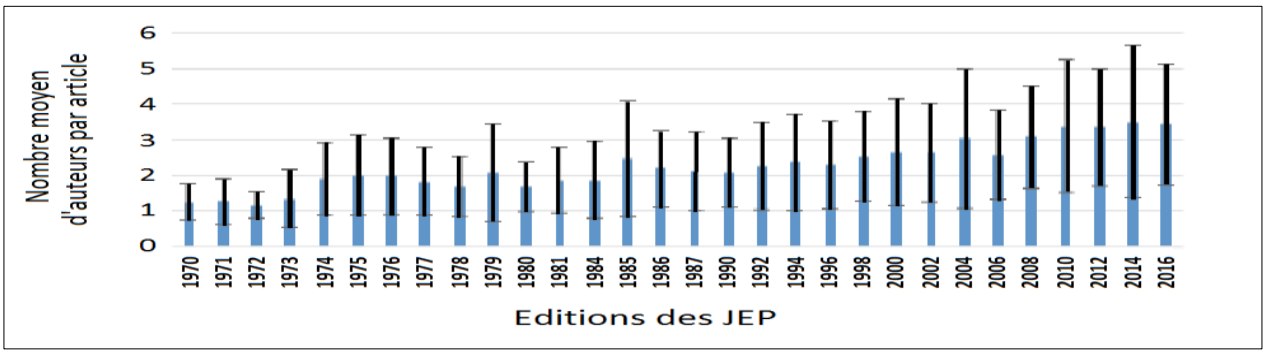



*Figure 2. Nombre moyen d'auteurs par article au fil des éditions des JEP.*

[4] Contrairement à nos attentes, nous n'avons pas constaté d'évolution notable de la proportion de laboratoires francophones (*vs*. laboratoires français) comme affiliation principale des premiers auteurs au cours des 5 éditions étudiées (environ 1/5). Depuis les années 2000, on observe pourtant une augmentation de premiers auteurs post-doctorants ayant fait leur thèse hors de France, mais ils sont majoritairement affiliés à un laboratoire français lorsqu'ils présentent une communication aux JEP.

### 3.4 Nombre moyen d'articles par auteur

La Figure 4 informe sur le nombre moyen d'articles parus dans les Actes des JEP par auteur (toutes éditions confondues). Les chercheurs n'ayant publié qu'un seul article sont majoritaires, étant donné qu'ils représentent près de 2/3 des noms de chercheurs présents dans les Actes. Quant aux chercheurs ayant entre 2 et 5 articles, ils constituent ¼ des publications. Si l'on s'intéresse aux chercheurs les plus prolifiques à travers le temps (Figure 5), on note les 50 articles produits par Louis-Jean Boë et les 41 travaux de Jean-Paul Haton en 31 éditions des JEP. D'autres chercheurs ont 20 articles ou plus : J.-L. Schwartz (33), J. Caelen (31), D. Fohr (30), G. Linares (30), J.-F. Bonastre (30), C. Abry (29), M. Adda-Decker (29), V. Aubergé (28), B. Teston (24), R. Sock (24), C. Benoit (23), P. Perrier (23), B. Guerin (22), F. Bechet (21), B. Harmegnies (20), G. Bailly (20), G. Perennou (20), H. Meloni (20), N. Vallée (20).

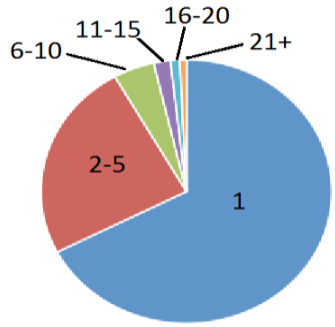



*Figure 4. Auteurs (proportion) : nombre total de publications dans les Actes des JEP.*

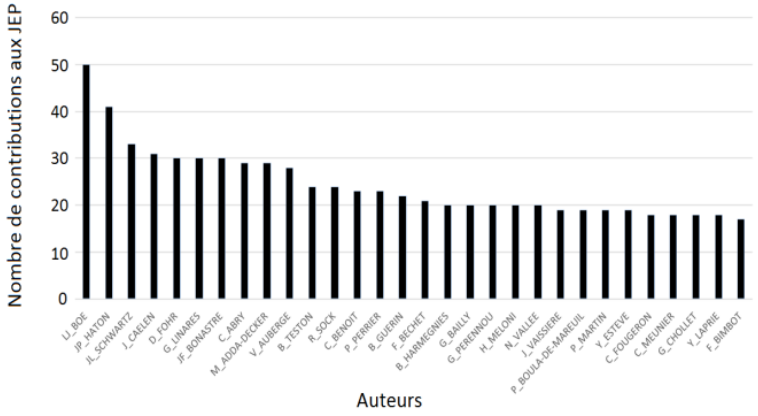



*Figure 5. Les 30 auteurs ayant le plus de contributions aux JEP (toutes éditions confondues).*

# 4 Etude textométrique des Actes des JEP

Un corpus de textes rassemblant l'ensemble des articles des JEP a été constitué. Pour cela, les documents les plus anciens ont d'abord été OCRisés puis corrigés avant d'être convertis au format brut. Après élimination de 8 articles en anglais et 15 articles présentant une proportion importante d'erreurs de numérisation, le corpus retenu compte 1997 articles. Celui-ci a alors été soumis au logiciel TXM pour une étude textométrique. Après lemmatisation (réalisée avec l'étiqueteur morphosyntaxique TreeTagger), il représente un volume d'environ 6 millions d'occurrences de mots, 140000 mots distincts et 120000 lemmes distincts. L'analyse des spécificités mise en œuvre par TXM repose sur une approche mesurant la distribution des mots dans différentes parties du corpus et basée sur la loi hypergéométrique (Lafon, 1980). Ici, les spécificités ont été calculées sur le corpus lemmatisé et partitionné par année afin de mettre en évidence les termes caractéristiques de chaque année. L'analyse a été réalisée sur l'ensemble des lemmes présentant une fréquence minimum de 200 et après élimination des mots grammaticaux (prépositions, pronoms, déterminants, conjonctions). Le

tableau 1 résume les résultats obtenus, avec les termes les plus significatifs regroupés par décennie. Il montre que les 30 premières années d'existence des JEP sont surtout marquées par la présence de termes renvoyant aux travaux menés en reconnaissance et en synthèse de la parole. Quant aux années 2000-2010, elles sont marquées par l'émergence de thématiques davantage liées aux SHS, tels que l'utilisation de corpus, la prosodie ou l'acquisition du langage, ainsi que par la mise en avant de travaux portant sur la production et la modélisation de la parole.

*Tableau 1. Spécificités lexicales associées à chaque décennie.*

| | |
|---|---|
| 1970 – 1979 | machine, appareil, vocal, synthétiseur, conduit, aire, fréquence, phonème, opérateur, filtre, automate, signal, linéaire, contour, fondamental, mélodie, synthèse |
| 1980 – 1989 | codeur, auditif, vérification, canal, intensité, spectre, vocodeur, analyseur, variabilité, phonétique, segment, distance, ton |
| 1990 – 1999 | réseau, fenêtre, vitesse, sémantique, modèle, module, représentations, dialogue, cible, système, apprentissage, probabilité, vecteur, détection, bruit |
| 2000 – 2009 | corpus, transcription, prosodie, modélisation, prosodique, voix, adaptation, enfant, codage, langue, production, ton |
| 2010 – 2016 | geste, production, F0, enfant, perception, données, automatique, classification, contour, pause |

# 5 Conclusion

L'histoire des Journées d'Etudes sur la Parole débute à la fin des années 60, c'est-à-dire bien après la création du Congrès International des Sciences Phonétiques. C'est sans doute l'une des raisons qui explique que le démarrage des JEP est marqué par une forte dominance des Sciences et Technologies de l'Information et de la Communication, qui sont alors en plein essor. Cette présence des STIC aux JEP sera toujours vérifiée, même si la part des chercheurs en SHS augmentera à travers le temps, pour être à peu près identique à celle des STIC de nos jours. Cette évolution des participants aux Journées d'Etudes sur la Parole est sans doute associée à l'évolution des thématiques abordées : si les premières éditions mettent davantage l'accent sur des thématiques reliées à la synthèse et/ou reconnaissance vocale, les derniers volets des JEP voient leurs Actes davantage marqués par des problématiques de sciences humaines.

La question qui se pose dorénavant est celle de l'avenir de ces journées scientifiques. Vont-elles réussir à se maintenir malgré le nombre croissant de manifestations scientifiques, notamment à l'international, et la diminution des ressources allouées aux (jeunes) chercheurs ? Les JEP vont-elles conserver ce savant équilibre entre chercheurs confirmés et étudiants, entre STIC et SHS, et bien sûr entre science et convivialité ? Vont-elles voir apparaitre de nouveaux acteurs issus de champs disciplinaires connexes (physiciens, neuroscientifiques, linguistes, cliniciens, didacticiens) ? En outre, l'un des défis pour les années à venir est de savoir si les JEP seront en mesure d'accueillir un nombre plus important de chercheurs issus de laboratoires situés hors de France. Le maintien, voire le développement de cette manifestation scientifique, pourrait passer par l'accueil d'un plus grand nombre de participants provenant de laboratoires plus diversifiés, à travers toute la francophonie.

# Références

BOË L.J., LIENARD J.S. (1988). La communication parlée est-elle une science ? Eléments de discussion et de réflexion suivis de repères chronologiques. Actes des *XVIIèmes Journées d'Etudes sur la Parole*, 79-92.

CARRE R. (1973). Allocution de Monsieur René Carré, Présidént du Groupe de la Communication Parlée, représentant le G.A.L.F. Actes des 4èmes Journées d'Etudes du Groupe de la Communication Parlée, 11-13.

GROSSETTI M. (1994). Sciences de la parole : genèse d'une communauté scientifique en France. Actes des *XVèmes Journées d'Etudes sur la Parole*, 3-10.

GUERIN-PACE F., SAINT-JULIEN T., LAU-BIGNON A.W. (2012). Une analyse lexicale des titres et mots-clés de 1972 à 2010. Revue *L'espace géographique*, 41, 4-30.

LAFON P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. Revue *Mots*, 1, 127-165.

STURM P. (2015). International Phonetic Congresses : the shift in research practices and areas of interest over 44 years. Actes des *44èmes International Congress of Phonetic Sciences*, https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0182.pdf

# Peut-on distinguer perceptivement huit accents régionaux en français parlé en Europe ? Une réponse à base de *crowdsourcing*

Mathieu Avanzi[1] & Philippe Boula de Mareüil[2]
(1) FNRS & VALIBEL, Univ. catholique de Louvain, Louvain-la-Neuve, Belgique
(2) LIMSI, CNRS & Univ. Paris-Saclay, Orsay, France
mathieu.avanzi@uclouvain.be ; philippe.boula.de.mareuil@limsi.fr

## RESUME

Nous présentons ici les résultats d'une enquête portant sur huit accents régionaux en français parlé en Europe : dans le nord, le sud-est et le sud-ouest de la France, en Alsace, Bretagne, Corse, Suisse et Belgique. Un test perceptif à grande échelle a été mené, à base de *crowdsourcing*, utilisant l'Internet pour recueillir les réponses de plus d'un millier de participants afin d'examiner si des auditeurs des régions ou pays susmentionnés sont capables de distinguer les accents correspondants. Les résultats suggèrent que ce sont plutôt trois grandes catégories d'accents (Nord-Ouest, Est et Sud) qui peuvent être discriminées.

## ABSTRACT

**Can eight regional accents be distinguished perceptually in European French? A crowdsourcing-based answer**

This paper presents the results of a survey addressing eight regional accents in European French: in the North, the South-East and the South-West of France, in Alsace, Brittany, Corsica, Switzerland and Belgium. A large-scale perception test was conducted, based on crowdsourcing, using the Internet to collect the responses of a thousand of participants, to examine whether listeners from the regions or countries mentioned are capable of distinguishing the corresponding accents. Results suggest that three broad categories of accents (North-West, South and East) can be distinguished.

MOTS-CLES : accents régionaux, géolinguistique, dialectologie perceptive, *crowdsourcing*
KEYWORDS: regional accents, geolinguistics, perceptual dialectology, crowdsourcing

# 1 Introduction

L'espace figure parmi les tout premiers facteurs pris en compte, tant par la dialectologie traditionnelle que par la sociolinguistique variationniste, pour expliquer l'hétérogénéité des pratiques langagières (Gadet, 2007 ; Mufwene & Vigouroux, 2012). Il nous donne des étiquettes pour dénommer aussi bien les langues (ou dialectes) que les accents, par exemple en français : ainsi va-t-on parler d'alsacien, de corse et d'accents corse ou alsacien en français, même si coexistent en Alsace des parlers alémaniques et franciques, en Corse des parlers d'origine toscane ou génoise. L'accent alsacien est-il distinct d'accents belges ou suisses ; l'accent corse est-il distinct d'accents continentaux du Midi de la France ? Les aires des dialectes et des accents ne coïncident pas nécessairement ; si les dialectes sont assez bien documentés et cartographiés depuis la fin du XIX$^{e}$ siècle (Gilliéron & Edmont, 1902–1910), les accents à l'intérieur de la langue


10.21437/JEP.2018-43

nationale (définis comme des ensembles de traits de prononciation liés à l'origine linguistique, géographique ou sociale des locuteurs d'une langue plus ou moins standardisée) n'ont suscité l'intérêt des linguistes que plus récemment.

Les accents régionaux, définis avant tout par la perception et les représentations qu'on en a, ont notamment été investis par la dialectologie perceptive (Preston, 1989 ; Innàccaro & Dell'Aquila, 2001, *inter alia*), dont un des objectifs est d'étudier dans quelle mesure différentes variétés d'une même langue peuvent être identifiées et catégorisées. Cette approche a été développée pour le français, avec des tests de perception incluant 5 points d'enquête en France (Woehrling & Boula de Mareüil, 2006), 3 en Belgique (Boula de Mareüil & Bardiaux, 2011), 4 en Suisse (Racine *et al*. 2013), 4 en Afrique de l'Ouest (Boula de Mareüil & Boutin, 2011), 8 au Québec (Remysen, 2016), mais chaque expérience était assez limitée, n'impliquant que quelques dizaines d'auditeurs. Une méthodologie à base de *crowdsourcing* (production participative à grande échelle) a depuis permis de passer à quelques centaines voire quelques milliers d'auditeurs, auxquels étaient présentés des accents français d'Amérique du Nord, d'Afrique et d'Europe (Boula de Mareüil *et al*., 2017), mais chaque variété était représentée par un seul locuteur.

D'autres expériences ont été conduites, centrées sur la France, la Suisse et la Belgique, touchant également une large audience, à travers des sites grand public ou les réseaux sociaux (Avanzi & Boula de Mareüil, 2017). Au total, des échantillons (d'une durée comparable aux tests précédents, soit une douzaine de secondes), provenant de 120 locuteurs (de 5 régions différentes dans chacun de ces pays) ont été présentés à des centaines d'auditeurs de ces mêmes pays. Dans une première série d'expériences, les auditeurs devaient identifier le pays d'origine des locuteurs : ils ont obtenu 60 % d'identification correcte en moyenne, avec des effets significatifs de l'origine des auditeurs, de l'âge, du statut socio-économique et de l'origine des locuteurs. Dans une deuxième série d'expériences, des auditeurs de France, de Suisse et de Belgique devaient identifier la région d'origine des locuteurs à l'intérieur de chacun de ces trois pays (avec un choix forcé entre 5 possibilités). Les résultats, quoiqu'au-dessus du niveau du hasard, se sont révélés moins bons, avec 31 % d'identification correcte en moyenne. Des analyses complémentaires ont été conduites pour évaluer le rôle de l'origine des auditeurs, de l'âge et de l'origine des locuteurs, ainsi que de leur interaction. Elles ont montré des patrons de réponses asymétriques entre les trois pays étudiés : la France (ou, en France, Paris), semble agir comme un pôle d'attraction et un catalyseur d'unification. Les locuteurs les plus jeunes, en particulier, sont plus souvent associés à sa façon de parler, quand leur accent n'est pas clairement identifiable. La Suisse, cependant, semble mieux résister que la Belgique à ce processus d'homogénéisation.

Dans l'étude que nous venons de résumer, toutefois, ne figurait ni l'accent alsacien, qui peut être confondu avec les accents belges et suisses (Woehrling, 2009), ni les accents méridionaux du sud-est et du sud-ouest de la France, ou encore de Corse. Dans cet article, nous nous proposons de réfléchir à une question en apparence simple : peut-on, pour le français parlé en Europe, distinguer les huit accents suivants : Nord, Bretagne, Alsace, Suisse, Belgique, Sud-Ouest, Sud-Est et Corse ? Comme tout dépend du *on*, il est nécessaire d'avoir un nombre suffisant d'auditeurs de chacune de ces régions (ou chacun de ces pays), d'où l'importance du *crowdsourcing*. Pour répondre à cette question, nous présenterons le corpus ainsi que la méthode employés (section 2) et les résultats obtenus, globalement et en fonction de l'origine des auditeurs (section 3), avant de conclure (section 4).

## 2 Corpus et méthode

Outre la Suisse et la Belgique, six régions ont été retenues dans cette étude : le Nord, la Bretagne à l'Ouest, l'Alsace à l'Est, le Sud-Ouest, le Sud-Est et la Corse. Des traits de prononciation, a priori, peuvent en effet distinguer le français parlé dans ces régions (Avanzi, 2017) : nous nous contenterons de citer le mot *moins*, dont la consonne finale tend à se faire entendre dans le Sud-Ouest. Pour la France, les enregistrements utilisés dans cette étude ont été collectés, sur le terrain, auprès de locuteurs qui, en plus du français, parlaient une langue régionale (picard, alsacien, breton ou gallo, occitan, corse). Souvent âgés, ces locuteurs étaient comme les Belges et les Suisses bien ancrés dans la région où ils ont été enregistrés, y ayant passé la plus grande partie de leur vie. Pour chacune des huit régions étudiées, nous avons sélectionné quatre locuteurs, à partir desquels nous avons extrait des segments de parole d'une durée comprise entre 10 et 15 secondes, en veillant que les échantillons ne contiennent aucun indice lexical orientant l'identification (des mots comme *septante* ou *nonante* entraînant un biais vers la Suisse ou la Belgique, par exemple), ni trop de disfluences (*euh* d'hésitation, répétitions, etc.). L'expérience d'identification a été mise au point avec le logiciel Qualtrics, lequel permet de faire des sondages en ligne, à partir de n'importe quel navigateur, sur ordinateur, téléphone portable multifonction ou tablette. Les participants, contactés via les réseaux universitaires et les réseaux sociaux, devaient cliquer sur un lien qui les amenait sur la plateforme d'enquête. Ils étaient informés qu'ils allaient entendre plusieurs extraits sonores, et qu'ils devaient identifier l'origine géographique des locuteurs. Pour ce faire, ils devaient appuyer sur le bouton du lecteur multimédia intégré, et cliquer, après écoute du stimulus, n'importe où à l'intérieur de la région colorée de la carte reproduite dans la Figure 1. Les participants pouvaient jouer l'extrait sonore autant de fois qu'ils le souhaitaient, mais ne pouvaient pas revenir en arrière après avoir validé leur réponse.

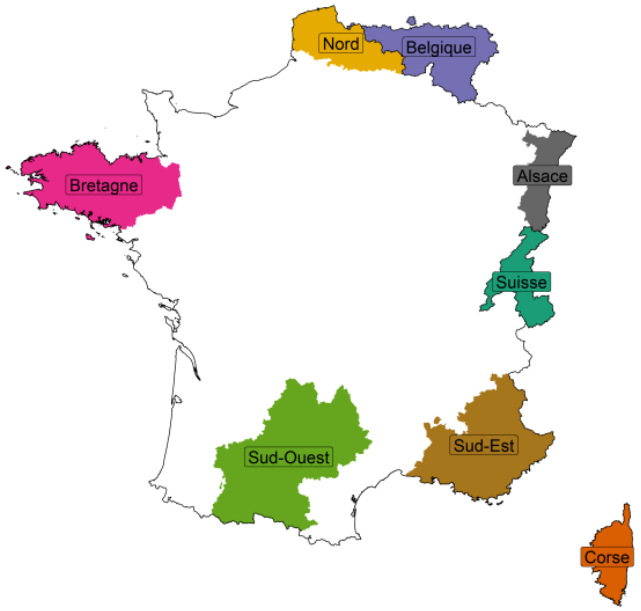



FIGURE 1: Carte présentant les régions et leurs labels proposée aux auditeurs lors de l'expérience.

Pour éviter que l'enquête ne soit trop longue, dans le but de recueillir les réponses d'un maximum de participants, seuls 16 extraits sonores parmi les 32 que comptait l'expérience étaient présentés à chaque sujet. Les stimuli étaient sélectionnés semi-aléatoirement, un algorithme permettant que

chacun des fichiers sons soit joué le même nombre de fois. À la fin de l'expérience, il était demandé aux participants d'indiquer leur âge, leur sexe, le pays et le code postal de la localité dans laquelle ils avaient passé la plus grande partie de leur jeunesse. Un champ libre leur permettait d'en dire plus quant à leur connaissance des accents régionaux. Au total, 1575 francophones ayant déclaré avoir passé la plus grande partie de leur jeunesse en France, Suisse ou Belgique ont pris part à cette expérience. Un des objectifs de cette étude étant de vérifier l'hypothèse selon laquelle on reconnaît mieux l'accent de sa propre région que celui des autres, nous avons codé l'origine géographique des participants dans notre base de données. À cette fin, nous avons considéré la Suisse et la Belgique comme des régions à part entière. Pour la France, nous avons eu recours à des tables de correspondance permettant de déterminer la région d'origine des participants à partir des codes postaux : pour l'Alsace, la Bretagne et la Corse, nous avons considéré les participants originaires des régions éponymes ; pour le Sud-Ouest, nous avons considéré les participants originaires de Midi-Pyrénées, pour le Nord ceux de Nord-Pas-de-Calais et pour le Sud-Est ceux de Provence-Alpes-Côte-D'azur. Les sujets se répartissent comme rapporté dans la Table 1.

| **Région** | **N** | **%** | **sexe** | | **âge** | | | |
|---|---|---|---|---|---|---|---|---|
| | | **du total** | **H** | **F** | **moyenne** | **min.** | **max.** | **écart type** |
| **Alsace** | 79 | 5 | 34 | 45 | 29,3 | 14 | 69 | 12,4 |
| **Belgique** | 118 | 7,5 | 54 | 64 | 28,7 | 17 | 80 | 14,4 |
| **Bretagne** | 145 | 9,2 | 69 | 76 | 29,8 | 14 | 89 | 12,3 |
| **Corse** | 10 | 0,6 | 2 | 8 | 35,4 | 20 | 70 | 18,0 |
| **Nord** | 63 | 4 | 39 | 24 | 30,6 | 15 | 71 | 11,6 |
| **Sud-Est** | 107 | 6,8 | 45 | 62 | 32,9 | 16 | 72 | 11,8 |
| **Sud-Ouest** | 102 | 6,5 | 45 | 57 | 31,2 | 14 | 67 | 12,5 |
| **Suisse** | 48 | 3 | 30 | 18 | 31,5 | 14 | 65 | 13,7 |
| **Autre** | 903 | 57,3 | 451 | 452 | 32,7 | 10 | 87 | 13 |
| **Total** | **1575** | **100** | **769** | **806** | **29,3** | **14,0** | **69,0** | **12,4** |

TABLE 1 : Participants à l'enquête, avec, leur région d'origine, leur nombre, le pourcentage par rapport au total, leur sexe et leur âge.

Il est intéressant de constater que l'échantillon est relativement bien équilibré en ce qui concerne le sexe des participants, et que malgré le format de l'expérience (accessible seulement en ligne et diffusée majoritairement sur les réseaux universitaires et sociaux), notre panel est relativement varié du point de vue de l'âge. Malgré tous nos efforts, nous ne pouvons que constater le nombre décevant de Corses, qui peut s'expliquer par la relativement faible population de l'île de beauté — la Corse du Sud, par exemple, est le deuxième département le moins peuplé de France après la Lozère. En Haute-Corse, de plus, nous n'avons pas souhaité mobiliser les réseaux militants auxquels les locuteurs appartenaient, pour ne pas biaiser les résultats par une reconnaissance de la voix plutôt que de l'accent.

# 3 Résultats

## 3.1 Statistiques

Les analyses statistiques ont été conduites avec le logiciel R version 3.3.2 (R Development Core Team, 2016). Pour vérifier si les variétés ont été identifiées au-dessus du niveau du hasard, nous avons réalisés différents tests *t* en fixant la valeur de *p* à 0,99. Pour examiner l'effet de l'origine

des locuteurs et/ou des auditeurs sur les scores d'identification obtenus, nous avons utilisé des modèles de régressions logistiques généralisés à effets mixtes, avec une fonction *logit* — R package *lme4* (Bates *et al.*, 2013) —, dans lesquels nous avons inclus les locuteurs et les auditeurs comme effets aléatoires. Les valeurs de *p* ont été obtenues à l'aide des fonctions *drop1* de la librairie *lme4* et *lsmeans* de la librairie éponyme. Enfin, pour visualiser dans un plan cartésien la distance relative entre les variétés perçues, nous avons eu recours à des techniques d'échelonnements multidimensionnels (MDS) de la librairie *MASS* (Venables & Ripley, 2002). Tous les graphiques ont été réalisés avec le package *ggplot2*.

## 3.2 Résultats globaux

Dans un premier temps, nous avons calculé le pourcentage de réponses obtenues pour chaque groupe de locuteurs, sans tenir compte de l'origine géographique des auditeurs. Nous avons ainsi pu obtenir la matrice de confusion de la Table 2.

| | | **origine prédite** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Alsace** | **Belgique** | **Bretagne** | **Corse** | **Nord** | **Sud-Est** | **Sud-Ouest** | **Suisse** |
| **origine réelle** | **Alsace** | **39,8** | 11,2 | 9,6 | 0,7 | 9,2 | 0,9 | 1,3 | 27,5 |
| | **Belgique** | 13,2 | **38,9** | 8,8 | 3,3 | 16,9 | 2,1 | 2,8 | 13,9 |
| | **Bretagne** | 11,2 | 6,3 | **45,9** | 2,7 | 18,1 | 2,8 | 4,0 | 8,9 |
| | **Corse** | 7,0 | 7,6 | 9,9 | **17,9** | 6,8 | 21,8 | 21,0 | 8,1 |
| | **Nord** | 10,8 | 5,8 | 37,2 | 2,0 | **29,2** | 4,2 | 3,3 | 7,4 |
| | **Sud-Est** | 1,5 | 1,1 | 2,7 | 11,7 | 1,4 | **38,4** | 42,0 | 1,2 |
| | **Sud-Ouest** | 2,0 | 0,4 | 3,8 | 15,2 | 1,4 | 33,8 | **42,3** | 1,1 |
| | **Suisse** | 11,2 | 8,3 | 14,5 | 3,9 | 13,9 | 7,0 | 6,5 | **34,9** |

TABLE 2 : Matrice de confusion, tous participants et toutes régions confondus (%).

En moyenne, avec 35,9 % d'identification correcte, les huit régions sont reconnues bien mieux qu'au hasard (12,5 %, tous les tests *t* donnant des valeurs de p<0,001). Pour vérifier si les différences entre les scores obtenus sont significatives, nous avons appliqué un modèle dans lequel la réponse (codée comme VRAI/FAUX) était la variable dépendante, la région des locuteurs (8 possibilités), la réponse des auditeurs (8 possibilités) et leur interaction étaient les prédicteurs, les stimuli et les auditeurs étant les variables aléatoires. Les résultats de ce test statistique ont révélé qu'il n'y a pas d'effet de région — les Corses, avec 17,9 % d'identification correcte, ne sont pas moins bien reconnus que les Bretons, avec 45,9 % d'identification correcte, pour ne prendre que les deux cas extrêmes — ni d'effet de réponse : les participants n'ont pas cliqué plus souvent sur la région Bretagne (16,1 %) que sur la région Corse (7,2 %), pour ne prendre de nouveau que les deux cas extrêmes. Cette absence d'effets simples est vraisemblablement due à la présence d'une interaction significative entre la région des locuteurs et la réponse des auditeurs ($\chi^2(7)=3769,1$ ; $p<0,0001$). Compte tenu de cette interaction significative, 8 modèles distincts ont été appliqués (un pour chaque région) afin de mettre au jour d'éventuelles différences entre les scores d'identification correcte à l'intérieur de chacune des huit régions. De ces modèles, il est ressorti que la variable « réponse des auditeurs » était toujours significative (tous les modèles mettent en évidence un effet significatif de réponse, avec des valeurs de $p<0,0001$). Pour rendre plus fluide la lecture, nous ne commenterons ici que les différences impliquant les pourcentages de réponses les plus élevés pour chaque région (en grisé dans la Table 2, les valeurs correctes sont en gras).

Pour l'Alsace, la Belgique, la Bretagne, le Sud-Ouest et la Suisse, les valeurs les plus hautes (celles en grisé dans la table 2) sont significativement différentes de toutes celles des lignes parentes, ce qui suggère par exemple que, pour les locuteurs alsaciens, la différence entre les réponses Alsace (39,8 %) et Suisse (27,5 %) n'est pas due au hasard. Pour la Corse, la différence n'est pas significative entre les réponses Corse (17,9 %) et Sud-Est d'une part (21,8 %), Corse et Sud-Ouest d'autre part (21,0 %). Pour le Nord, le pourcentage de bonnes réponses (29,2 %) est significativement inférieur ($p$<0,0001) au pourcentage de réponses Bretagne (37,2 %). Enfin, pour le Sud-Est, il n'y a pas de différence significative entre le pourcentage de bonne réponse (38,4 %) et le pourcentage de réponses Sud-Ouest (42 %). En résumé, il ressort qu'à l'intérieur de chaque région, le score d'identification correcte est significativement différent de tous les autres scores à l'intérieur de cette même région, sauf pour la Corse (les Corses étant confondus avec des locuteurs du Sud-Est et du Sud-Ouest), pour le Nord (où c'est la réponse Bretagne qui arrive en tête des suffrages) et pour le Sud-Est (dont les locuteurs sont confondus avec ceux du Sud-Ouest).

Sur le plan perceptif, se dégagent trois groupes de variétés, que l'échelonnement multidimensionnel (non représenté ici par manque de place) permet de visualiser. Les trois variétés du Sud-Ouest, du Sud-Est et de Corse forment un premier groupe : elles ont souvent été confondues entre elles, la différence entre la Corse et les deux régions du sud du continent étant plus importante que celle entre Sud-Est et Sud-Ouest. D'autre part, on trouve les variétés de l'est de la francophonie d'Europe (la Belgique étant perceptivement équidistante de la Suisse et de l'Alsace) et les variétés du nord-ouest de la France (Nord et Bretagne).

## 3.3 Effet de l'origine des auditeurs

Dans un second temps, nous avons examiné l'effet de l'origine des auditeurs sur l'identification d'accents régionaux, notre hypothèse étant qu'un accent devrait être mieux reconnu par des locaux, qui en sont familiers. Pour plus de clarté, nous avons isolé les réponses obtenues pour les locuteurs de chacune des huit régions et avons comparé les scores des auditeurs habitant ces régions à ceux qui n'y habitaient pas : 8 modèles linéaires généralisés avec la réponse VRAI/FAUX comme variable dépendante, l'interaction avec l'origine des auditeurs (locaux ou non-locaux) et leur réponse (8 possibilités), les stimuli et les auditeurs étant entrés comme variables aléatoires. Les scores d'identification correcte des locaux et des autres, pour chacune des huit régions, sont données dans la Table 3. Toutes les différences sont significatives ($p$<0,001, hormis pour la Corse, où aucun test statistique n'a pu être mené, compte tenu du trop faible nombre de participants — les valeurs sont donc données à titre indicatif. Nous avons ensuite réalisé un MDS sur la base des réponses des locaux uniquement, pour chacune des huit régions, ce qui nous a permis de visualiser aisément les confusions commises par les auditeurs.

On voit dans la Table 3 que les locaux sont meilleurs que les autres pour reconnaître l'accent de leur région. Si l'on compare à présent les 8 MDS présentés dans la Figure 2, on constate que certains groupes d'auditeurs ne confondent jamais l'accent de leur région avec celui des autres. Tel est le cas des Alsaciens, des Belges, des Corses et des Suisses, alors que d'autres groupes d'auditeurs ont plus de peine à distinguer l'accent de leur région de celui de régions proches : tel est le cas des Bretons et des participants du Nord, mais également des participants du Sud.

| | Alsace | Belgique | Bretagne | Corse | Nord | Sud-Est | Sud-Ouest | Suisse |
|---|---|---|---|---|---|---|---|---|
| **locaux** | 78,8 | 75,3 | 58,5 | 93,1 | 55,0 | 51,4 | 64,9 | 53,7 |
| **non-locaux** | 37,6 | 35,9 | 44,7 | 17,4 | 28,1 | 37,4 | 40,1 | 34,2 |

TABLE 3 : Taux d'identification correcte de chaque région, selon l'origine des auditeurs (%).

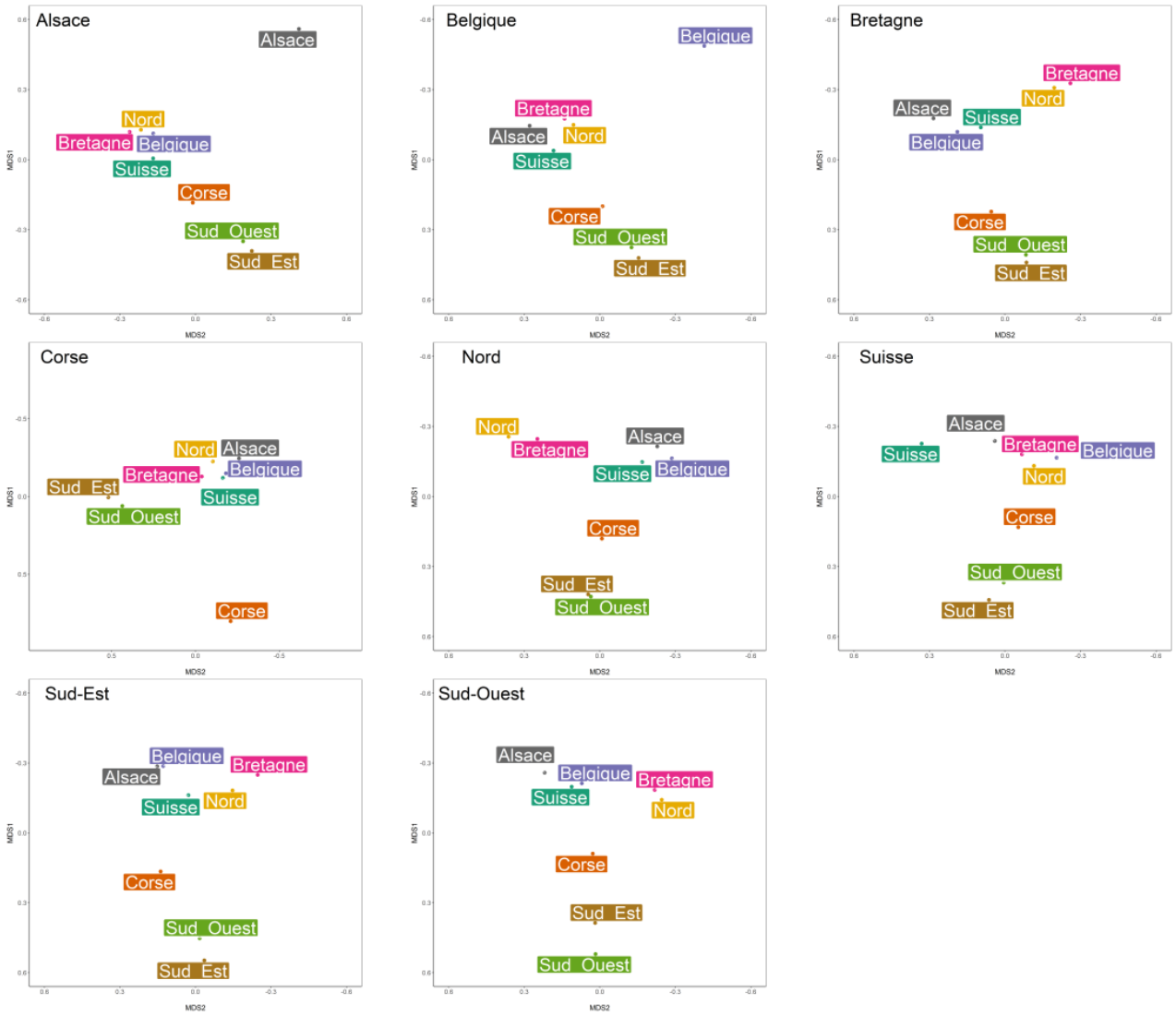



FIGURE 2 : Echelonnements multidimensionnels pour les huit régions étudiées, d'après les réponses obtenues par les locaux (leur origine étant à chaque fois indiquée en haut à gauche).

## 3.4 Zoom sur les régions du sud de la France

Les résultats présentés dans la section 3.2 ont permis de mettre le doigt sur des confusions constantes entre le Sud-Est et le Sud-Ouest ; ceux de la section 3.3 ont révélé que ces confusions sont également faites par les participants de ces régions. Dans cette dernière partie, nous avons cherché à évaluer et comparer les capacités des auditeurs du sud de la France à distinguer les locuteurs du Sud-Est de ceux du Sud-Ouest.

| Auditeurs du Sud-Est | Sud-Est | Sud-Ouest |
|---|---|---|
| **Sud-Est** | **51,4** | 30,2 |
| **Sud-Ouest** | 37,7 | **43,5** |

| Auditeurs du Sud-Ouest | Sud-Est | Sud-Ouest |
|---|---|---|
| **Sud-Est** | **27,7** | 47,1 |
| **Sud-Ouest** | 23,9 | **64,9** |

TABLE 4 : Matrices de confusion simplifiées pour les auditeurs du Sud-Est et du Sud-Ouest (%).

Un modèle linéaire généralisé a été appliqué, avec comme variable dépendante la réponse (VRAI/FAUX), l'origine des participants (Sud-Est *vs* Sud-Ouest), l'origine des locuteurs (Sud-Est *vs* Sud-Ouest), la réponse sélectionnée (8 possibilités) et toutes les interactions simples (les stimuli et les auditeurs étant entrés comme variables aléatoires). Il a montré une absence d'effets simples : les auditeurs du Sud-Est ne sont pas plus performants que les auditeurs du Sud-Ouest ; les locuteurs du Sud-Est ne sont pas mieux identifiés que ceux du Sud-Ouest, mais il y a une interaction significative entre l'origine et la réponse des auditeurs ($\chi^2(7)=71,692$ ; $p<0,0001$), ainsi qu'une interaction significative entre l'origine des locuteurs et la réponse des auditeurs ($\chi^2(7)=40,123$ ; $p<0,0001$). En résumé, il ressort que les auditeurs du Sud-Ouest sont plus performants que ceux du Sud-Est quand ils identifient des locuteurs du Sud-Ouest ($p<0,0001$) : comparer 64,9 % avec 43,5 %. A contrario, les auditeurs du Sud-Est sont meilleurs quand il s'agit d'identifier leur propre variété régionale en comparaison à celle du Sud-Ouest ($p<0,01$) : comparer 51,4 % avec 27,7 %. Par ailleurs, on constate que les auditeurs du Sud-Est ne sont pas significativement meilleurs lorsqu'il s'agit de faire la part entre les locuteurs du Sud-Ouest et ceux du Sud-Est : comparer 51,4 % avec 43,5 %. A contrario, les auditeurs du Sud-Ouest sont clairement meilleurs quand il s'agit de reconnaître leur propre variété à côté de celle du Sud-Est ($p<0,001$) : comparer 64,9 % avec 27,7 %.

# 4 Conclusion

Pour reprendre la question posée dans le titre, « peut-on perceptivement identifier des accents d'Alsace, de Belgique, de Bretagne, de Corse, du Sud-Est, du Sud-Ouest et de Suisse, », la réponse se doit d'être nuancée. Globalement, ce sont plutôt trois grands groupes qui se distinguent : Nord-Est (incluant la Suisse et la Belgique), Nord-Ouest et Sud (incluant la Corse). Si ce résultat confirme des expériences antérieures (Woehrling, 2009, *inter alia*) et les généralise, le nombre de participants que le *crowdsourcing* nous a permis de recruter apporte un éclairage nouveau, même s'il pouvait être attendu : il est notable que des locaux parviennent, parmi huit possibilités, à reconnaître l'accent de leur région à plus de 50 %. En dépit du grand nombre d'auditeurs qui ont pris part à ce test d'identification, ce travail met également en lumière le fait que le *crowdsourcing* n'est pas la panacée : le nombre de Corses reste insuffisant, suggérant que cette approche ne doit pas se substituer à des enquêtes de terrain. Au contraire, les deux démarches sont complémentaires et peuvent se féconder l'une l'autre. Dans tous les cas, une analyse des stimuli est à mener, de même qu'une analyse des résultats par locuteur, afin de cerner si des prototypes se dégagent pour les différents accents.

# Remerciements

Mathieu Avanzi a reçu le financement du Fonds National de la Recherche Scientifique (subside n° 24901170). Ce travail s'inscrit également dans les activités du projet « Donnez votre français à la science » (DFS) du programme « Langues et Numérique » 2016 de la Délégation Générale à la Langue Française et aux Langues de France (DGLFLF). Nous remercions les locuteurs et les nombreux auditeurs qui ont rendu possible ce travail.

# Références


AVANZI, M. (2017). *Atlas du français de nos régions*. Paris : Armand Colin.

AVANZI, M. & BOULA DE MAREÜIL, P. (2017). Identification of regional French accents in (northern) France, Belgium and Switzerland. *Journal of Linguistic Geography*, 5/1, 17–40.

BATES, D. M., MÄCHLER, M., BOLKER, B. & WALKER, S. (2013). lme4: Linear mixed-effects models using Eigen and S4. *R package*.

BOULA DE MAREÜIL, P. & AKISSI BOUTIN, B. (2011). Évaluation et identification perceptives d'accents ouest-africains en français. *Journal of French Language Studies,* 21, 361–379.

BOULA DE MAREÜIL, P. & BARDIAUX, A. (2011). *Perception of French, Belgian and Swiss accents by French and Belgian Listeners. 4th ISCA Tutorial and Research Workshop on Experimental Linguistics*. Paris 47–50.

BOULA DE MAREÜIL, P., SCHERRER, Y. & GOLDMAN, J.-P. (2017). Combien d'accents en français ? Focus sur la France, la Belgique et la Suisse. *Bulletin suisse de linguistique appliquée,* 104, 91–103.

GADET, F. (2007). *La variation sociale en français*. Paris/Gap : Ophrys.

GILLIÉRON, J. & EDMONT, E. (1902–1910). *Atlas linguistique de la France (ALF)*. Paris : Champion.

INNÀCCARO, G. & DELL'AQUILA, V. (2001). Mapping languages from inside: notes on perceptual dialectology. *Social and Cultural Geography*, 2/3, 265–280.

MUFWENE, S. & VIGOUROUX, C. (2012). Individuals, populations, and timespace: Perspectives on the ecology of language *Cahiers de linguistique*, 38/2, 111–138.

PRESTON, D. R. (1989). *Perceptual dialectology*. Dordrecht : Foris.

RACINE, I., SCHWAB, S. & DETEY, S. (2013). Accent(s) suisse(s) ou standard(s) suisse(s) ? Approche perceptive dans quatre régions de Suisse romande. In A. Falkert (dir.), *La perception des accents du français hors de France*. Mons : CIPA, 41–59.

R DEVELOPMENT CORE TEAM. (2016). R: A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

REMYSEN, W. (2016). Langue et espace au Québec : les Québécois perçoivent-ils des accents régionaux ? In D. Gavinelli & C. Molinari (Eds.), *Lingue, culture, mediazioni, (Espaces réels et imaginaires au Québec et en Acadie : enjeux culturels, linguistiques et géographique)*. Milan : LED, 31–57.

VENABLES, B. & RIPLEY, B. D. (2002). *Modern Applied Statistics with S*. New York : Springer.

WOERHLING, C. (2009). *Accents régionaux en français. Perception, analyse et modélisation à partir de grands corpus.* Thèse de doctorat de l'Uiversité Paris Sud, Orsay.

WOERHLING, C. & BOULA DE MAREÜIL, P. (2006). Identification d'accents régionaux en français : perception et analyse. *Revue Parole,* 37, 25–65.

# L'information accentuelle est-elle représentée dans le lexique mental des locuteurs du français ?

Amandine Michelas, Sophie Dufour
Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France

michelas@lpl-aix.fr, sophie.dufour@lpl-aix.fr

## RESUME

Une particularité prosodique du français est que l'accent primaire ne permet pas de distinguer deux mots de sens différents. A l'aide d'un paradigme d'amorçage de répétition, nous avons examiné si l'information accentuelle est représentée dans le lexique mental des français. En comparaison à une condition contrôle, nous avons observé une diminution des temps de réaction, que le mot cible « ban'**deau** » accentué sur sa syllabe finale soit précédé de l'amorce « ban'**deau** » elle-même accentuée sur sa syllabe finale ou qu'il soit précédé de l'amorce « bandeau » inaccentuée. Un tel résultat suggère que les mots sont stockés indépendamment de la présence ou non d'un accent en français. D'un point de vue théorique plus général, ce résultat est compatible avec les modèles abstractionnistes de la reconnaissance des mots parlés dans lesquels les informations acoustiques non pertinentes à l'identification des mots seraient écartées du signal de parole avant l'accès au lexique.

## ABSTRACT

**Is stress information represented in the mental lexicon of French speakers?**

One striking characteristic of French is that stress does not allow distinguishing two words of different meanings. Using the repetition priming paradigm, we examined whether stress information is represented in the mental lexicon of French speakers. In comparison to a control condition, we observed shorter reaction times both when the target word "ban'**deau**" stressed on its final syllable was preceded by the prime "ban'**deau**" also stressed on its final syllable and when it was preceded by the word "bandeau" unstressed. Such a result suggests that words are stored independently of the presence or absence of stress in French. At a more theoretical level, this result is compatible with abstract models of spoken word recognition in which acoustic details irrelevant to identification are discarded from the speech signal before lexical access.

MOTS-CLES : Perception de la parole, reconnaissance des mots, représentations abstraites, prosodie du français, accent primaire

KEYWORDS: Speech perception, word recognition, abstract representations, French prosody, stress.



10.21437/JEP.2018-44

# 1 Introduction

Une des caractéristiques prosodiques les plus frappantes du français par rapport aux autres langues romanes est le fait que l'accent primaire ne permet pas de distinguer deux mots de sens différents comme c'est le cas en espagnol (ex. '**be**be « elle boit » vs. be'**be** « bébé »). Au contraire, en français l'accent primaire[1] affecte une unité plus grande que le mot, le syntagme accentuel, et c'est toujours la dernière syllabe du syntagme accentuel qui porte l'accent. Cet accent est obligatoire et est caractérisé par deux corrélats acoustiques principaux : l'allongement de la durée de la dernière syllabe du syntagme et une montée de fréquence fondamentale (f0) associée à cette syllabe lorsque le syntagme n'est pas en position finale d'énoncé (Jun & Fougeron, 2000 ; Welby, 2006). Le fait qu'en français l'accent affecte toujours la dernière syllabe du syntagme accentuel a une conséquence importante : un même mot peut-être accentué ou non en en fonction de sa position au sein du syntagme. Par exemple, le mot BANDEAU est accentué dans la phrase « On m'a parlé [d'un petit BAN'**DEAU**]$_{SA}$ » car il est situé en position finale du syntagme accentuel ($_{SA}$). Au contraire, ce même mot n'est pas accentué dans la phrase « On m'a parlé [d'un bandeau '**rouge**]$_{SA}$ » car il n'est pas en position finale du syntagme accentuel. De ce fait, même si l'accent ne permet pas de créer de contrastes de sens au niveau du mot, les français sont confrontés quotidiennement aux versions accentuées et inaccentuées des mots. Dans cette étude, nous nous sommes donc demandé si l'accent est intégré ou non aux représentations mentales des mots chez les français comme cela a été montré chez les auditeurs espagnols (Soto-Faraco et al., 2001).

Au regard des modèles théoriques de la reconnaissance des mots parlés, seuls les modèles exemplaristes dans lesquels chaque mot est associé à de multiples *tokens* encodant des détails acoustiques fins (Goldinger, 1998) postulent que l'information accentuelle est stockée dans le lexique mental des français. Au contraire, les modèles abstractionnistes (McClelland & Elman, 1986), supposant une première étape de normalisation du signal de parole au cours de laquelle les détails acoustiques fins non pertinents pour l'identification des mots seraient écartés du signal de parole, postulent que la présence ou non d'un accent primaire chez les locuteurs du français ne serait pas intégrée à leurs représentations lexicales.

Dans la présente étude, nous avons cherché à départager ces deux types de modèles en utilisant un paradigme d'amorçage de répétition à long-terme. Ce paradigme consiste à présenter dans un premier temps un bloc de mots (bloc amorce) aux participants sur lesquels ils doivent réaliser une tâche (i.e., décision lexicale). Dans un second temps, un second bloc de mots (bloc cible) leur est présenté, la moitié des mots ayant déjà été rencontrée dans le premier bloc, l'autre n'ayant jamais été rencontrée. Typiquement, les mots qui sont répétés sont reconnus plus rapidement que les mots non répétés. L'atténuation de cet effet d'amorçage de répétition lors de la modification d'une dimension particulière entre les deux blocs (par exemple lors de la modification de la présence/absence d'un accent), indiquerait en accord avec les modèles à exemplaires que le même mot prononcé avec une différence d'accent activerait différentes représentations lexicales et que des spécificités liées à la présence ou non d'un accent primaire seraient stockées en mémoire. Au contraire, aucune modulation de l'effet d'amorçage en fonction de la présence/absence d'un accent

[1] Notons qu'il existe également un accent secondaire en français, qui est optionnel, qui affecte le début du syntagme accentuel et qui n'a pas les mêmes corrélats acoustiques que l'accent primaire. Contrairement à l'accent primaire, l'accent secondaire n'est pas accompagné d'un allongement de la syllabe. Il est seulement associé à une montée de f0 qui affect le début du premier mot de contenu du syntagme accentuel (Welby, 2006).

indiquerait en accord avec les théories abstractionnistes de la reconnaissance des mots parlés que le même mot prononcé avec des différences accentuelles activerait la même représentation lexicale. Des participants de langue maternelle française ont eu à réaliser une tâche de décision lexicale, dans laquelle ils devaient décider le plus rapidement et le plus précisément possible si les stimuli présentés constituaient ou non un mot de la langue française. De façon à nous assurer de l'origine lexicale de l'effet d'amorçage de répétition susceptible d'être observé sur les mots, les pseudo-mots pouvaient être répétés soit avec une accentuation identique ou soit avec une accentuation différente. La logique sous-jacente à une telle manipulation est que des effets d'origine lexicale devraient être plus importants pour les mots que pour les pseudo-mots, ces derniers par définition n'étant pas associés à une représentation lexicale.

# 2 Méthode

## 2.1 Participants

48 participants de langue maternelle française ont pris part à l'expérience. Tous les participants étaient des étudiants d'Aix-Marseille Université et ont tous rapporté n'avoir aucun trouble de l'audition, de la parole ou trouble neurologique.

## 2.2 Matériel

48 mots cibles bisyllabiques de structure CVCV ont été sélectionnés sur la base de données Lexique.org (New et al., 2005) et ont été utilisés à la fois comme amorces et comme cibles. La fréquence moyenne de ces mots était de 7,60 par million d'occurrences. De plus, 48 pseudo-mots cibles bisyllabiques de structure CVCV, servant également comme amorces, ont été créés en changeant le dernier phonème de mots réels (ex. « baisi » créé à partir du mot « baiser »). 16 autres mots utilisés comme amorces contrôles ont été également sélectionnés et 16 autres pseudo-mots utilisés également comme amorces contrôles ont été créés.

Afin d'obtenir les versions accentuées et inaccentuées de chacun des stimuli, nous avons demandé à une locutrice de langue maternelle française de prononcer les 64 mots et 64 pseudo-mots au sein de phrases porteuses dans lesquelles ils pouvaient être accentués ou non en fonction de leur position dans la phrase (ex. On m'avait parlé [d'un bandeau **'bleu]$_{SA}$** qui était joli vs. On m'avait parlé [d'un petit ban**'deau]$_{SA}$** qui était joli ; ex. On m'avait parlé [d'un baisi **'bleu]$_{SA}$** qui était joli vs. On m'avait parlé [d'un petit bai**'si]$_{SA}$** qui était joli ). Afin d'éviter les effets de coarticulation lié à la parole en contexte, chaque mot a d'abord été extrait de sa phrase porteuse, puis présenté auditivement à la locutrice. Celle-ci devait répéter chacun des mots et des pseudo-mots dans leurs versions accentuées et inaccentuées. Les 128 répétitions ainsi obtenues ont été enregistrées à une fréquence d'échantillonnage de 44 100 Hz, segmentées puis normalisées en intensité à un niveau de 70 dB SPL.

Des analyses acoustiques ont été ensuite conduites grâce au logiciel Praat (Boersma & Weenink, 2015) sur ces 128 répétitions afin de s'assurer que celles-ci étaient produites avec les patrons accentuels attendus. La figure 1 illustre le mot « bandeau » et le pseudo-mot « baisi » produits dans leurs versions accentuées et inaccentuées.

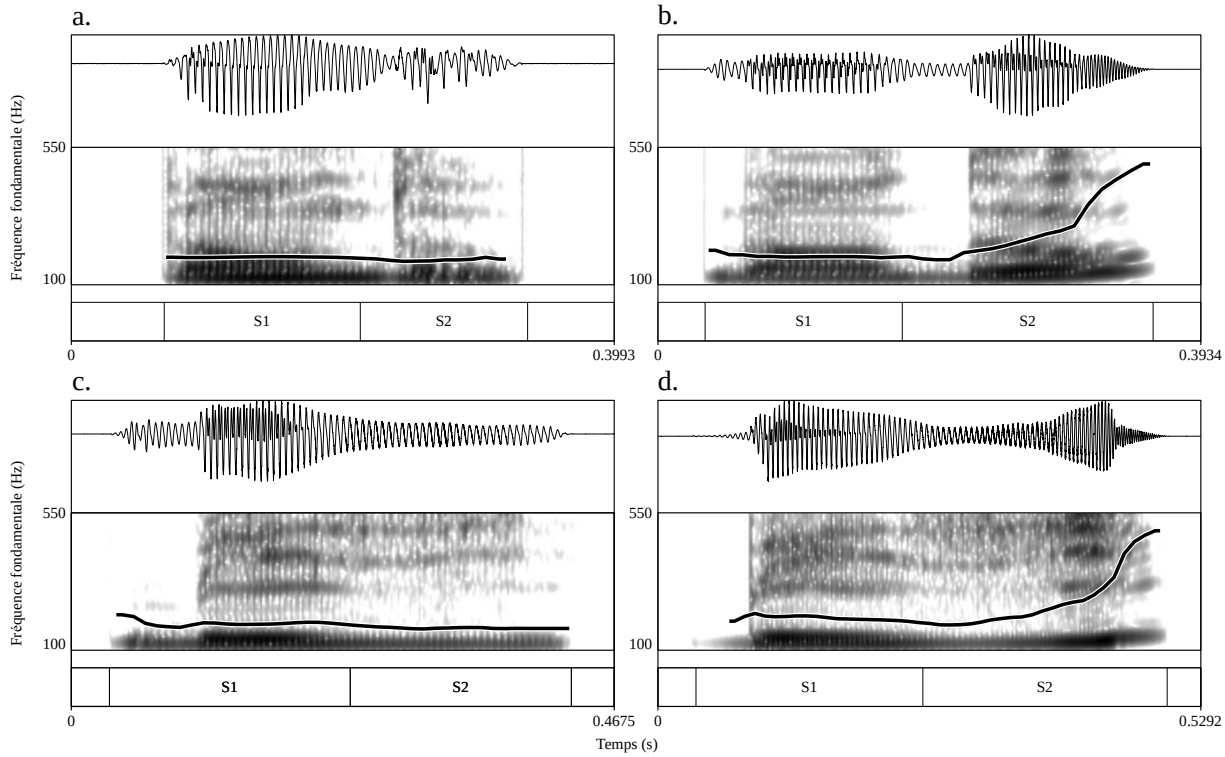



FIGURE 1. Onde sonore, spectrogramme et courbe de f0 pour le mot « bandeau » dans sa version inaccentuée (a) et accentuée (b) et le pseudo-mot « baisi » dans sa version inaccentuée (c) et accentuée (d).

L'allongement de la syllabe finale de chaque stimulus ainsi que la montée de f0 associée à cette syllabe finale ont été mesurés (cf. Table 1).

| | | Version inaccentuée | | Version accentuée | |
|---|---|---|---|---|---|
| | | 1ère syllable du mot | 2ème syllable du mot | 1ère syllable du mot | 2ème syllable du mot |
| Mots | Durée syllabique (ms) | 144 | 143 | 142 | 239 |
| | Minimum de f0* (Hz) | 187 | 176 | 182 | 171 |
| | Maximum de f0* (Hz) | 181 | 174 | 178 | 402 |
| | Montée de f0 (%) | -3 | -1 | -2 | 136 |
| Pseudo-mots | Durée syllabique (ms) | 153 | 153 | 153 | 242 |
| | Minimum de f0* (Hz) | 185 | 180 | 184 | 180 |
| | Maximum de f0* (Hz) | 183 | 178 | 183 | 374 |
| | Montée de f0 (%) | -1 | -1 | -1 | 107 |

TABLE 1. Caractéristiques acoustiques des mots et pseudo-mots cibles dans leurs versions accentuées et inaccentués. * Pour les versions inaccentuées des mots et pseudo-mots, les valeurs minimum et maximum de f0 correspondent aux valeurs de début et de fin du plateau de f0 observé.

Comme attendu, les 48 mots cibles ainsi que les 48 pseudo-mots cibles avaient une syllabe finale plus longue ($t(47)=17.73$, $p<.0001$) et plus haute que la première syllabe ($t(47)=21.26$, $p<.0001$) uniquement dans leurs versions accentuées. Les mêmes vérifications ont été effectuées pour les 16 mots et les 16 pseudo-mots utilisés comme amorce contrôle.

Deux blocs de stimuli ont déjà été créés. L'un était utilisé comme bloc amorce et l'autre comme bloc cible. A l'intérieur de chaque bloc, la moitié des stimuli était accentuée et l'autre moitié était inaccentuée. Le bloc cible était constitué des 48 mots cibles et des 48 pseudo-mots cibles. Parmi les mots et les pseudo-mots cibles, 16 ont été utilisés dans la condition « répétée accentuation identique», 16 dans la condition « répétée accentuation différente » et 16 dans la condition contrôle (i.e. non répétée). Le bloc amorce était également constitué de 48 mots et de 48 pseudo-mots. Parmi les mots et les pseudo-mots, 16 consistaient en la répétition des cibles avec accentuation identique, 16 en la répétition des cibles avec une accentuation différente et les 16 autres correspondaient aux amorces contrôles.

Afin que chaque cible (mots, pseudo-mots) soit entendue dans chaque condition d'amorçage (répétée accentuation identique, répétée accentuation différente, contrôle) et que chaque participant n'entende qu'une seule fois le même mot cible, trois listes expérimentales ont été créées. Les trois listes ont ensuite été divisées en deux sous-listes de manière à ce que chaque stimulus soit entendu dans sa version accentuée et non accentuée.

### 2.3 Procédure

Les participants, munis d'un casque audio, ont été testés individuellement dans une chambre insonorisée et les stimuli leur étaient présentés à un niveau sonore confortable. La présentation des stimuli était contrôlée par un ordinateur grâce au logiciel E-Prime (version 2.0, Psychology Software Tools). Pour chaque stimulus du bloc amorce et du bloc cible, les participants devaient indiquer le plus rapidement et le plus précisément possible s'il s'agissait d'un mot de la langue française ou non en fournissant la réponse « mot » avec leur main dominante. Les temps de réponse (TRs) étaient enregistrés à partir du début des stimuli. A l'intérieur de chaque bloc, les stimuli étaient présentés dans un ordre aléatoire. La réponse du participant et le début de présentation du stimulus suivant étaient séparés par un délai de 2 secondes. Les participants ont été testés sur une seule des sous-listes expérimentales et ont commencé l'expérience avec 12 essais d'entraînement.

## 3 Résultats

Uniquement les temps de réaction obtenus dans le bloc cible ont été analysés. Six items engendrant des taux d'erreurs supérieurs à 40% ont été exclus des analyses. Une donnée aberrante correspondant à un TR de 372ms ainsi que 25 données aberrantes correspondants à des TRs supérieurs à 2000ms ont été supprimées. Les analyses ont donc portées sur 3973 données. Les temps de réaction (TRs) moyens ainsi que les pourcentages de réponses correctes sont présentés dans la Figure 2.

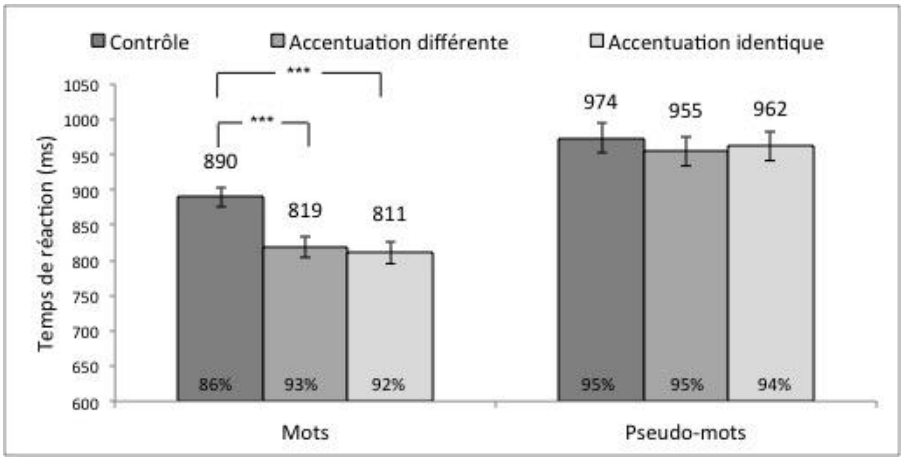



FIGURE 2. Temps de réaction moyen (en ms) et pourcentage de réponses correctes obtenus dans le bloc cible en fonction de la lexicalité et du type d'amorce (les barres représentent les erreurs standards).

Afin d'analyser ces temps de réaction, nous avons utilisé un modèle linéaire à effets mixtes sur les transformations logarithmiques des TRs (lme4 package in R-studio statistics Version 0.99.903). Ce modèle avait pour effets fixes le type d'amorce (contrôle/répétée accentuation identique/répétée accentuation différente) et la lexicalité (mot/pseudo-mot). La structure aléatoire du modèle incluait des pentes différentes pour les participants pour le facteur lexicalité et pour le facteur type d'amorce, des pentes différentes pour les mots pour le facteur type d'amorce ainsi que des interceptes différents pour les participants et les mots (Barr et al., 2013). Le modèle a révélé un effet principal significatif du type d'amorce ($F=34.27$, $p<.0001$) ainsi que de la lexicalité ($F=59.02$, $p<.0001$). L'interaction type d'amorce * lexicalité était également significative ($F=17.52$, $p<.0001$). Afin de décomposer cette interaction, le design 2x3 a été converti en un design 1x6. Les comparaison multiples ont été obtenues grâce à la fonction glht du package multcomp (Bretz et al., 2011) avec des p-values ajustées avec une correction de Bonferroni.

Les temps de réponse étaient plus rapides pour les mots cibles précédés d'une amorces répétée avec accentuation identique (811ms) que pour les mots cibles précédés d'une amorce contrôle (890ms ; $z=-8.82$ ; $p< .0001$). Ils étaient également plus rapides pour les mots cibles précédées d'une amorce répétée avec accentuation différente (819ms) que pour les mots cibles précédés d'une amorce contrôle (890ms ; $z=-7.72$, $p<.0001$). Au contraire, aucune différence n'a été observée entre les mots cibles précédés d'une amorce répétée avec accentuation identique (811ms) et les mots cibles précédées d'une amorce répétée avec accentuation différente (819ms ; $z=1.13$ ; $p>.20$). Nous avons également observé aucune différence entre les pseudo-mots quelque soit le type d'amorce qui était présenté (pseudo-mots accentuation identique vs. pseudo-mots contrôles : $z=-1.15$, $p>.20$ ; pseudo-mots accentuation différente vs. pseudo-mots contrôles : $z=-2.16$, $p>.20$; pseudo-mots accentuation identique vs. pseudo-mots accentuation différente: $z=-1.00$, $p>.20$).

Les taux de réponses correctes ont été analysés à l'aide d'un modèle logit à effets mixtes. Ce modèle avait pour effets fixes le type d'amorce (contrôle/répétée accentuation identique/répétée accentuation différente) et la lexicalité (mot/pseudo-mots). La structure aléatoire du modèle incluait un intercepte différent pour les participants et les mots mais pas de pentes différentes car l'inclusion de celles-ci ne permettait pas au modèle de converger (Barr et al., 2013). Les effets globaux ont été obtenus grâce à la fonction afex::mixed (Sinngmann et al., 2013). Le modèle a révélé un effet principal significatif de la lexicalité ($X^2=7.35$, $p<.01$). L'effet du type d'amorce échouait à atteindre

la significativité ($X^2$=5.60, p=.06), alors que l'interaction type d'amorce * lexicalité était significatif ($X^2$=9.43, p<.01). Afin de décomposer cette interaction, le design 2x3 a été converti en un design 1x6. Les comparaison multiples ont été obtenues grâce à la fonction glht du package multcomp (Bretz et al., 2011) avec des p-values ajustées avec une correction de Bonferroni. Le modèle a révélé plus de réponses correctes pour les mots cibles précédés d'une amorce répétée avec accentuation identique (92%) que pour les mots cibles précédés d'une amorce contrôle (86% ; z=-3.31 ; p< .05). Les participants ont également donné plus de réponses correctes lorsque les mots cibles étaient précédés d'une amorce répétée avec accentuation différente (93%) que lorsque les mots cibles étaient précédés d'une amorce contrôle (86% ; z=-3.83, p<.01). Au contraire, aucune différence de performance n'a été observée entre les mots cibles précédés d'une amorce répétée avec accentuation identique (92%) et les mots cibles précédés d'une amorce répétée avec accentuation différente (93% ; z=-0.56 ; p>.20). Aucune différence significative n'a été observée pour les pseudo-mots quelque soit le type d'amorce qui était présenté (p>.20).

# 4 Discussion

L'hypothèse sous-jacente à notre recherche était que si les indices liés à la présence/absence d'un accent sont intégrés au lexique mental des locuteurs du français, une atténuation de l'effet d'amorçage de répétition devrait être observée lorsque des amorces et des cibles diffèrent du point de vue de leur accentuation. Au contraire, si les indices liés à la présence/absence d'un accent ne sont pas intégrés aux représentations lexicales des locuteurs du français, aucune modulation dans l'effet d'amorçage de répétition ne devrait être observée lorsque des amorces et des cibles diffèrent sur leur accentuation. Les résultats obtenus dans cette étude ne montrent aucune diminution de la taille de l'effet d'amorçage de répétition lors d'un changement d'accentuation entre les amorces et les cibles. En effet, les temps de réponse des participants étaient plus rapides pour les mots cibles précédés d'une amorce répétée avec accentuation identique que pour les mots cibles précédés d'une amorce contrôle. Les temps de réponse étaient également plus rapides pour les mots cibles précédés d'une amorce répétée avec accentuation différente que pour les mots cibles précédés d'une amorce contrôle. Cependant, de façon cruciale, aucune différence de temps de réaction n'a été observée entre les mots cibles précédés d'une amorce répétée avec accentuation identique et les mots cibles précédés d'une amorce répétée avec accentuation différente. Une telle observation plaide en faveur de l'existence de représentations n'encodant pas l'information accentuelle chez les locuteurs du français et indique qu'un même mot prononcé avec ou sans accent active la même représentation lexicale de base.

L'avantage majeur du paradigme d'amorçage de répétition à long-terme que nous avons utilisé est qu'il nous permet de sonder la nature des représentations lexicales et plus particulièrement quel type d'information elles sont susceptibles d'encoder. L'origine lexicale de nos effets est attestée par l'absence d'effet d'amorçage de répétition avec des pseudo-mots. Les pseudo-mots, n'étant par définition associés à aucune représentation lexicale, ne sont pas réactivés à un niveau lexical de traitement lors de leur seconde présentation et aucun effet d'amorçage de répétition n'a ainsi été observé. Dans deux études précédentes (Michelas et al. 2016 ; Michelas et al., 2017) et à l'aide d'une tâche « même-different », nous avons montré que les auditeurs du français étaient capables de traiter la différence entre des mot accentués et des mots inaccentués et ceci à un niveau plus abstrait de traitement que de simples distinctions acoustiques contrairement à ce qui avait été suggéré dans des précédentes études (Dupoux et al., 1997). Toutefois même si les français sont capables de percevoir les différences entre un mot inaccentué et accentué, la présente étude montre que l'information accentuelle n'est en aucun cas encodé dans leurs représentations lexicales.

D'un point de vue théorique plus général, un tel résultat s'intègre dans le débat actuel concernant le format des représentations lexicales et en particulier sur l'inclusion ou non de variations accentuelles dans ces représentations. Comme nous l'avons vu précédemment, seuls les modèles exemplaristes de la reconnaissance des mots parlés postulent que l'information accentuelle est intégrée aux représentations lexicales des locuteurs du français. Néanmoins, une façon de rendre compte de nos résultats dans un cadre exemplariste est d'envisager que toutes les variations acoustiques ne sont pas encodées dans les représentations lexicales et qu'un poids plus important serait accordé à certaines variations. Toutefois, le fait que nos manipulations de durée et de f0 n'aient pas eu d'impact sur la reconnaissance des mots semble être difficilement réconciliable avec une vision purement exemplariste de la reconnaissance des mots parlés. Nos résultats semblent donc être d'avantage en accord avec les modèles abstractionnistes tels que TRACE (McClelland & Elman, 1986) dans lesquels les détails acoustiques non pertinents pour identifier les mots en français sont écartés lors d'une première phase de normalisation du signal de parole et dans lesquels des représentations symboliques abstraites n'encodant pas de variation acoustique sont contactées.

Pour conclure, il apparaît donc que l'information accentuelle soit traitée et stockée différemment selon la structure prosodique des langues. Pour une langue telle que le français où l'information accentuelle n'est pas pertinente pour identifier deux mots de sens différents, les détails acoustiques liés à la présence ou non d'un accent seraient utilisés à un niveau pré-lexical de traitement (ex. pour segmenter mots ; Christophe et al., 2004) mais ne seraient pas intégrés aux représentations lexicales. Au contraire, pour des langues à accent lexical tel que l'anglais ou l'espagnol, les représentations mentales des mots incluraient des paramètres accentuels tels que la position de l'accent (Soto-Farraco et al., 2001).

# Remerciements

Cette recherche, menée dans le cadre du Labex BLRI (ANR-11-LABX-0036) et l'Institut Convergence ILCB (ANR-16-CONV-0002), a reçu le soutien du gouvernement français, par le biais de l'Agence nationale de la Recherche (ANR) et l'Initiative d'Excellence d'Aix-Marseille Université (A*MIDEX).

# Références

BARR, D. J., LEVY, R., SCHEEPERS, C., TILY, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language* 68(3), 255-278.

BOERSMA, P., WEENINK, D. (2015). *Praat. Doing phonetics by computer* (Version 5.4.01, 2015). Computer program: www.praat.org (Last viewed April 30, 2015).

BRETZ, F., HOTHORN, T., WESTFALl, P. H. (2011). Multiple comparisons using R. Boca Raton: CRC Press.

CHRISTOPHE, A., PEPERKAMP, S., PALLIER, C., BLOCK, E., MEHLER, J. (2004). Phonological phrase boundaries constrain lexical access I. Adult data. *Journal of memory and language* 51(4), 523-547.

DUPOUX, E., PALLIER, C., SEBASTIAN, N., MEHLER, J. (1997). A destressing "deafness" in French?. *Journal of Memory and Language* 36(3), 406-421.

GOLDINGER, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. Psychological review, 105(2), 251.

JUN, S. A., FOUGERON, C. (2000). A phonological model of French intonation. In A. Botinis (Eds), *Intonation* (pp. 209-242). Springer, Dordrecht.

MCCLELLAND, J. L., ELMAN, J. L. (1986). The TRACE model of speech perception. *Cognitive psychology* 18(1), 1-86.

MICHELAS, A., FRAUENFELDER, U. H., SCHÖN, D., DUFOUR, S. (2016). How deaf are French speakers to stress?. *The Journal of the Acoustical Society of America* 139(3), 1333-1342.

MICHELAS, A., ESTEVE-GIBERT, N., DUFOUR, S. (2017). On French listeners' ability to use stress during spoken word processing. *Journal of Cognitive Psychology*, 1-9.

SINGMANN, H., BOLKER, B., WESTFALL, J., AUST, F. (2015). *afex: Analysis of factorial experiments*. R package version 0.13–145.

SOTO-FARACO, S., SEBASTIAN-GALLES, N., & CUTLER, A. (2001). Segmental and suprasegmental mismatch in lexical access. *Journal of Memory and Language* 45, 412-432.

WELBY, P. (2006). French intonational structure: Evidence from tonal alignment. *Journal of Phonetic*s 34(3), 343-371.

# Évaluations perceptive et automatique de l'intelligibilité de la parole dégradée par simulation de la surdité professionnelle

Imed Laaridh[1] Julien Tardieu[2] Cynthia Magnen[2] Pascal Gaillard[3]
Jérôme Farinas[1] Julien Pinquier[1]
(1) IRIT, Université de Toulouse, CNRS, Toulouse, France
(2) MSHS-T (USR 3414), Université de Toulouse et CNRS, France
(3) CLLE (UMR 5263) Université de Toulouse et CNRS, France
prenom.nom@irit.fr[1], prenom.nom@univ-tlse2.fr[2,3]

## RÉSUMÉ

Cet article présente une étude comparative entre des mesures perceptive et automatique pour l'évaluation de l'intelligibilité de la parole dans des conditions dégradées. Il fait suite à une étude précédente qui a permis de proposer une méthodologie pour la simulation et l'évaluation automatique de l'effet de la presbyacousie sur l'intelligibilité de la parole. Nous proposons, dans ce travail, d'adapter cette approche à une autre pathologie de l'audition : la surdité par traumatisme en milieu professionnel. Dans ce cadre, un corpus de parole a été constitué afin de refléter différents niveaux de surdité professionnelle et a été soumis à des évaluations perceptive et automatique. L'évaluation perceptive a démontré des effets similaires des dégradations sur les locuteurs homme, femme et enfant. De plus, les mesures d'intelligibilité automatiques, fondées sur la reconnaissance automatique de la parole, sont très corrélées aux performances humaines.

## ABSTRACT

**Perceptual and automatic evaluations of the intelligibility of speech degraded by noise induced hearing loss simulation**

This study aims at comparing perceptual and automatic intelligibility measures on degraded speech. It follows a previous study that proposed an approach for the simulation and automatic evaluation of the effect of age-related hearing loss (presbycusis) on speech intelligibility. In this work, we propose to adapt this approach to a different hearing disorder : noise induced hearing loss (hearing loss caused by trauma in the professional context). Thus, a speech corpora simulating different levels of noise induced hearing loss is proposed and used to study perceptual and automatic intelligibility measures. The perceptual evaluation showed similar effects of the hearing disorder on the male, female and child speakers intelligibility. Also, the automatic intelligibility measure (based on automatic speech recognition scores) is proved to well represent the effects of the different severity levels of the hearing disorder. Indeed, high correlation measure are computed between the automatic and perceptual intelligibility measures.

MOTS-CLÉS : Mesure de l'intelligibilité de la parole, reconnaissance automatique de la parole, pathologies de l'audition, simulation de la surdité professionnelle.

KEYWORDS: speech intelligibility metric, automatic speech recognition, hearing disorders, noise induced hearing loss simulation.


10.21437/JEP.2018-45

# 1 Introduction

Plus de 3 millions de travailleurs en France sont exposés sur leur lieu de travail, d'un manière prolongée, à des niveaux de bruit potentiellement nocifs. En effet, et selon le centre d'information sur le bruit (CidB), environ 1200 cas de surdité professionnelle dus à ce type de traumatisme sonore sont diagnostiqués chaque année. Cette pathologie de l'audition est caractérisée par la présence de zones cochléaires mortes (i.e. détérioration de certaines parties des cellules cillées empêchant le codage tonotopyque). L'apparition de ces zones mortes s'accompagne de phénomènes d'acouphènes et d'hyperacousie résultant en des problèmes de compréhension de la parole à la fois dans les conditions de silence et de bruit.

Plus spécifiquement, cette pathologie se manifeste par une perte neuro-sensorielle dans les hautes fréquences du signal et par une encoche localisée dans la région autour de 4 kHz (McBride & Williams, 2001). La gravité de cette perte et de ce trou auditif dépend de la sévérité du traumatisme et peut évoluer dans le temps affectant d'autres bandes fréquentielles (notamment les bandes conversationnelles) avoisinantes les régions les plus touchées initialement. Ces troubles, lorsqu'ils ne sont pas soignés, peuvent compromettre les capacités de communication des personnes atteintes et engendrer des effets néfastes sur leurs vies personnelle et professionnelle tels que l'isolement (Strawbridge *et al.*, 2000) et la dépression (Gopinath *et al.*, 2009). Une des principales solutions aux pathologies auditives est l'utilisation d'audioprothèses. Ces outils permettent, en amplifiant certaines bandes fréquentielles, de donner une meilleure audibilité aux patients atteints. Cependant, comme rapporté dans (Vestergaard Knudsen *et al.*, 2010), environ 40% des patients équipés de ce type d'outils ne l'utilisent jamais (ou rarement). Ce rejet peut être expliquer, entre autres, par le peu de réglages spécifiques de ces prothèses à chaque utilisateur.

Dans la pratique clinique, l'évaluation des troubles auditifs se fait à l'aide d'audiogrammes tonaux ainsi que des tests vocaux. Les épreuves les plus utilisées sont celles de transcription du contenu linguistique de mots/phrases par les patients. L'aptitude d'un auditeur à reconnaître ce qui a été prononcé peut alors être utilisée comme une mesure de sa capacité auditive. Ces tests peuvent être utilisés dans différentes perspectives ; que ce soit dans la phase du diagnostique ou lors du réglage (et/ou de l'adaptation) des audioprothèses à chaque auditeur afin d'obtenir la meilleure possible intelligibilité de la parole.

Cependant, plusieurs limites peuvent êtres associées à ces épreuves perceptives. En effet, elles sont souvent très coûteuses en temps, notamment à cause de l'implémentation et du traitement souvent manuels des résultats de l'évaluation. De plus, et afin d'avoir une évaluation assez robuste, l'utilisation d'un jury d'auditeurs et non d'un seul est conseillée. Toutes ces caractéristiques limitent leur application clinique, notamment lors des phases de réglage spécifique de l'audioprothèse pour chaque utilisateur. Par ailleurs, le matériel linguistique utilisé dans ces tests est souvent assez limité (exemple : liste de mots de Fournier (Fournier, 1951)). Les patients sont alors rapidement familiers avec le contenu des tests. Cela, associé au fait que ces évaluations doivent souvent être répétées plusieurs fois afin de bien adapter l'audioprothèse dans différentes conditions sonores, peut ainsi compromettre leurs résultats (Hustad & Cahill, 2003).

Le reste de cet article est organisé comme suit. Dans la section 2, le contexte et les objectifs de ce projet sont expliqués. La section 3 décrit la méthodologie utilisée dans ce travail, notamment le processus de simulation des dégradations liées à la surdité professionnelle ainsi que le système de Reconnaissance Automatique de la Parole (RAP) utilisé. La section 4 présente les différents résultats des évaluations perceptive et automatique alors que la section 5 fournit quelques conclusions et

directions pour de futurs travaux.

# 2 Contexte et objectifs

Cette étude fait partie d'un projet pluridisciplinaire auquel participent différents acteurs du traitement du langage et de la parole : des informaticiens, des orthophonistes et des ORL. Ce projet fait suite à un travail précédent qui a permis de proposer et valider une méthodologie pour la simulation et l'évaluation automatique des troubles de l'audition liés à la presbyacousie sur l'intelligibilité de la parole (Fontan *et al.*, 2017).

Le but de ce projet est d'adapter cette méthodologie à d'autres troubles de l'audition, spécifiquement la surdité professionnelle. L'approche proposée permettra alors d'évaluer automatiquement (sur la base des résultats de la RAP) les effets de la surdité professionnelle sur l'intelligibilité et donc de mieux affiner et faciliter le réglage des audioprothèses proposé aux patients par les experts. Plus largement, le projet visera aussi à adapter cette méthodologie à d'autres langues que le Français (Anglais) et à d'autres conditions de dégradation de la parole (parole dans un milieu bruité). Dans un autre contexte, la méthodologie proposée pourra être utilisée pour l'évaluation objective de l'intelligibilité de la parole, la rééducation et le suivi clinique de l'évolution de la condition des patients atteints de troubles de la parole : dysarthrie, cancers de la cavité buccale et du pharynx, etc.

# 3 Méthodologie

La méthodologie proposée pour la simulation et l'évaluation de la surdité professionnelle consiste en 3 phases représentées dans la figure 1.

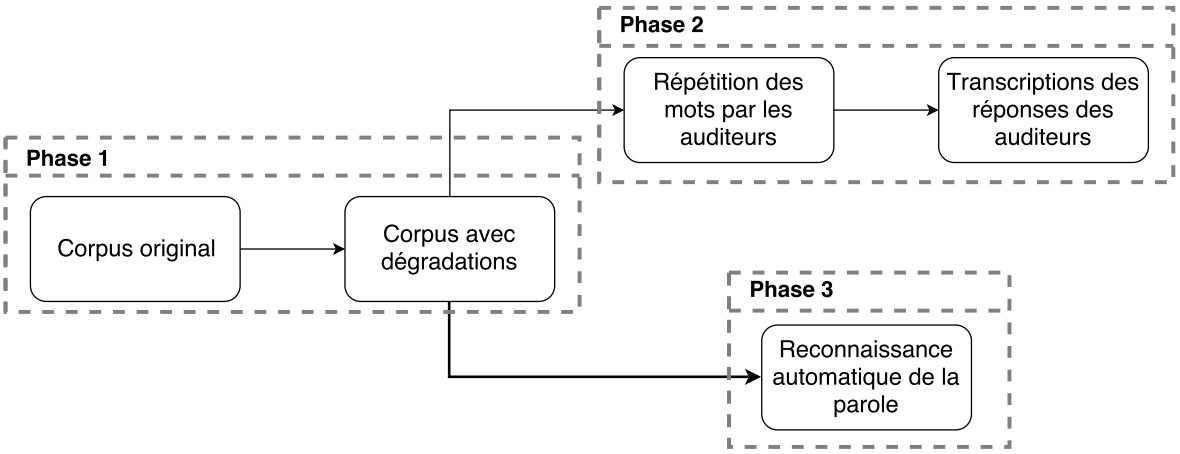



FIGURE 1: Diagramme des différentes phases de la méthodologie proposée.

La première phase décrit la constitution du corpus de données utilisé. Des enregistrements de mots issus de la liste de Fournier sont réalisés par trois locuteurs (un homme, une femme et un enfant).

Ensuite, différents niveaux de dégradation associés à la surdité professionnelle sont simulés sur les enregistrements afin de construire le corpus de parole utilisé lors des évaluations perceptive et automatique. La deuxième phase consiste en une évaluation perceptive de l'intelligibilité des enregistrements de la parole dégradés par 31 auditeurs. Ce protocole est décrit dans la sous-section 3.2. Les mêmes enregistrements sont aussi utilisés dans la troisième phase dans une tâche de RAP afin de mesurer des scores d'intelligibilité automatiques. Le système de RAP utilisé est décrit dans la sous-section 3.3.

## 3.1 Phase 1 : préparation de données

Le matériel linguistique utilisé dans ce protocole consiste en 60 mots (6 listes de 10 mots) issus des listes de mots proposées par Fournier en 1951 pour la mesure perceptive de l'intelligibilité (Fournier, 1951). Ces listes sont largement utilisées par les audioprothésistes pour l'évaluation de l'audition des patients. Tous les mots débutent avec une consonne et sont proposés sous la forme : article + nom (par exemple : « le parfum »).

Trois locuteurs : un homme (46 ans), une femme (47 ans) et une jeune fille (12 ans) ont participé à la production des enregistrements des mots. Les trois locuteurs avaient le Français comme langue maternelle et ont tous produit tous les mots de la liste. Les enregistrements ont été effectués dans une cabine audiométrique (PETRA [1]) à l'aide d'un microphone omnidirectionel Sennheiser MD46, d'une console de mixage TASCAM DM-3200 et d'un ordinateur MacPro équipé du logiciel Reaper.

Pour simuler la présence d'un bruit ambiant, nous avons choisi d'utiliser un bruit de type « brouhaha » construit à partir des enregistrements dans le silence en suivant la procédure décrite dans (Fontan *et al.*, 2017). Ce fichier de bruit a ensuite été mixé aux mots avec un rapport signal sur bruit de 5 dB. Les stimuli de parole avec et sans bruit ont ensuite été dégradés pour simuler les effets de la surdité professionnelle. La simulation a été effectuée dans MATLAB à partir des algorithmes initialement développés par (Nejime & Moore, 1997). À partir de données issues du terrain [2], 10 audiogrammes ont été simulés (voir tableau 1) allant d'une audition normale (niveau 1) à une surdité professionnelle grave (niveau 10).

TABLE 1: Audiogrammes utilisés dans la simulation de la surdité professionnelle, exprimés en pertes auditives (dB) pour chaque bande de fréquence (kHz).

| Niveau | .125 | .25 | .5 | .75 | 1 | 1.5 | 2 | 3 | 4 | 6 | 8 | 10 | 12 | 14 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 1 | 3 | 5 | 7 | 10 | 20 | 10 | 7 | 5 | 3 | 1 | 0 |
| 3 | 0 | 0 | 5 | 5 | 5 | 8 | 10 | 15 | 35 | 15 | 10 | 8 | 5 | 5 | 5 |
| 4 | 0 | 0 | 5 | 6 | 7 | 14 | 18 | 27 | 47 | 30 | 20 | 14 | 7 | 6 | 5 |
| 5 | 0 | 0 | 5 | 7 | 8 | 19 | 27 | 38 | 58 | 45 | 30 | 19 | 8 | 7 | 5 |
| 6 | 0 | 0 | 5 | 8 | 10 | 25 | 35 | 50 | 70 | 60 | 40 | 25 | 10 | 8 | 5 |
| 7 | 4 | 5 | 14 | 19 | 25 | 39 | 49 | 61 | 78 | 68 | 50 | 36 | 20 | 16 | 9 |
| 8 | 8 | 10 | 23 | 29 | 40 | 53 | 63 | 73 | 85 | 75 | 60 | 48 | 30 | 24 | 13 |
| 9 | 11 | 15 | 31 | 40 | 55 | 66 | 76 | 84 | 93 | 83 | 70 | 59 | 40 | 32 | 16 |
| 10 | 15 | 20 | 40 | 50 | 70 | 80 | 90 | 95 | 100 | 90 | 80 | 70 | 50 | 40 | 20 |

1. http://petra.univ-tlse2.fr
2. https://www.uvmt.org/sections.php?op=printpage&artid=568

## 3.2 Phase 2 : évaluation perceptive

L'évaluation perceptive consiste en une tâche d'écoute et de répétition des différents stimuli présentés dans les différentes conditions sonores. Après un entraînement sur 10 mots, chaque participant entendait les 60 mots de la liste auxquels étaient appliquées aléatoirement les 60 conditions du plan expérimental : dégradation(10)*locuteur(3)*bruit(2). La tâche du participant consistait à répéter oralement chaque mot entendu avant de passer au suivant. Les participants étaient assis à un mètre des deux haut-parleurs (Focal Solo 6 BE) dans la cabine audiométrique PETRA, et devant le microphone utilisé pour enregistrer les réponses. Le niveau sonore de diffusion des mots non dégradés était de 60 dBA (niveau 1 dans le silence). Les réponses ont ensuite été transcrites par les expérimentateurs afin de calculer les scores d'intelligibilité. Ce protocole d'évaluation a été retenu afin de forcer les auditeurs à fournir une réponse, éliminer les effets potentiels de leurs capacités orthographiques et n'impliquer que les composants de production et de perception de la parole dans l'évaluation.

31 participants (nombre de femmes = 20, âge moyen = 20,5 et écart type = 1,8) ont été sélectionnés pour participer à cette expérience. Les critères d'inclusion sont les suivants : francophones natifs (langue maternelle française, ayant toujours vécu en France), âgés de 18 à 30 ans inclus, étudiants dans des disciplines autres que musique, sciences du langage, langues étrangères ou psychologie, sans problèmes de vue non corrigés par des lentilles ou lunettes. Tous les participants ont été soumis à un audiogramme avec le logiciel AudioConsole afin de contrôler leur niveau d'audition.

## 3.3 Phase 3 : approche d'évaluation automatique

Dans ce travail, nous avons utilisés un système de RAP basé sur l'outil Sphinx-3 (Seymore *et al.*, 1998) distribué par Carnegie Mellon University (CMU). Il est important de noter qu'à l'inverse des applications classiques de la RAP, le but de ce travail n'est pas de proposer un système capable de reconnaître la parole dans des conditions dégradées et d'améliorer au plus ses performances en terme de taux d'erreur de mots reconnus (Word Error Rate - WER). L'objectif de ce travail est de proposer une approche automatique capable de simuler et de reproduire le comportement de la perception humaine face aux distorsions liées à la surdité professionnelle.

Les modèles acoustiques utilisés dans ce travail sont proposés par le Laboratoire d'Informatique de l'Université du Maine (LIUM) (Deléglise *et al.*, 2005) et appris sur plusieurs heures d'enregistrements radiophoniques issues du corpus ESTER (Galliano *et al.*, 2009). Il s'agit de modèles contextuels composés de 35 phonèmes et 5 types de pauses/silences (résultant en 5725 phonèmes en contexte) dont chaque état est représenté par un mélange de 22 lois gaussiennes (Gaussian Mixture Model - GMM). Chaque enregistrement est échantillonné à 16 kHz et une paramétrisation PLP (Hermansky, 1990) est utilisée. Puisque les modèles ont été appris sur de la parole produite uniquement par des hommes, le comportement du système de RAP n'était pas adapté à la parole produite par les locuteurs femme et enfant. Afin de pallier à ce problème, une adaptation de la longueur théorique du tract vocal (Vocal Tract Length Normalization - VTLN) a été utilisée (Wegmann *et al.*, 1996). Cette approche se fonde sur l'hypothèse de l'existence d'une relation linéaire entre la longueur du conduit vocal d'un locuteur et les fréquences des zones formantiques qui lui sont associées.

Deux modèles de langage ont été utilisés. Un premier modèle trigrammes de base, appelé BM, appris sur le corpus ESTER2 sur la base d'un lexique d'environ 62000 mots. Un second modèle bigrammes reflétant la composition syntaxique particulière de la liste de mots utilisé (article + nom). Ce modèle, appelé ML, a été appris sur des mots dissyllabiques commençant avec une consonne seulement

(environ 15000). Les fréquences utilisées pour les différentes formes du modèle sont celles définit dans (New *et al.*, 2007) et sont basées sur les sous-titres des films de la base de données Lexique 3.8.

# 4 Résultats

## 4.1 Évaluation de l'intelligibilité perceptive

Pour chaque mot, le score est égal à 1 si le mot est totalement reconnu (l'article précédent le mot n'est pas pris en compte) ou 0 s'il n'est que partiellement/ou non reconnu. Les scores d'intelligibilité correspondent ensuite aux pourcentages de mots reconnus pour chaque dégradation, chaque locuteur, dans le silence et dans le bruit. Les résultats sont présentés sur la figure 2. Les données ont été analysées en utilisant un modèle linéaire mixte généralisé, et ont révélé un effet significatif de la dégradation ($p<0{,}001$) et du bruit ($p<0{,}001$), mais un effet non significatif du locuteur (et donc de son âge et genre) ($p=0{,}63$). Ces résultats sont cohérents avec ceux obtenus dans nos travaux précédents sur l'effet de la presbyacousie simulée sur l'intelligibilité de mots (Fontan *et al.*, 2017).

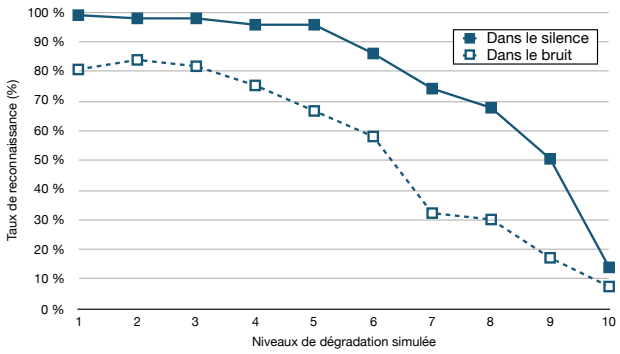


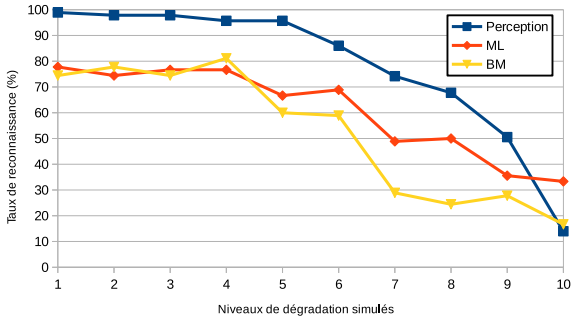

FIGURE 2: Scores d'intelligibilité perceptive dans le silence et dans le bruit.

## 4.2 Évaluation d'intelligibilité automatique

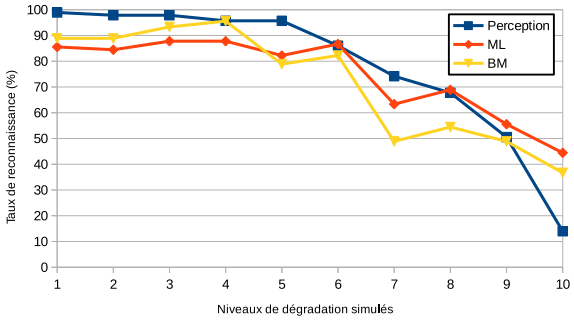
Nous avons appliqué l'approche de mesure automatique d'intelligibilité de la parole uniquement sur les stimuli en condition de silence. Ce choix vise à valider l'intérêt de l'approche avant de la confronter aux conditions les plus difficiles. Ce choix permet aussi une meilleure comparaison avec les résultats obtenus lors notre étude précédente (Fontan *et al.*, 2017).

La figure 3 rapporte les taux de reconnaissances automatique et perceptive moyens sur les trois locuteurs pour les différents niveaux de dégradation simulés. Deux stratégies de reconnaissance automatique sont retenues ; la première considère qu'un mot est reconnu automatiquement que s'il se retrouve en première position dans les propositions du système de RAP. La deuxième, indique si le mot cible fait partie des 10 mots les plus probables pour le système. Nous observons qu'à l'exception de la condition de dégradation la plus sévère 10, le taux de reconnaissance automatique est inférieur à celui observé chez les auditeurs humains. En effet, le taux de reconnaissance moyen atteint 77,8% pour les auditeurs humains contre 60,9% et 52,4% pour le système automatique en utilisant les modèles de langage ML et BM respectivement et la première stratégie de reconnaissance automatique. La capacité du système automatique, contrairement à la perception humaine, de compenser les dégradations les plus sévères (traduite par un taux de reconnaissance supérieur) nécessite une investigation plus approfondie. Sans surprise, nous observons que le modèle de langage ML, plus adapté aux listes de

mots utilisées dans notre protocole expérimental, atteint des taux de reconnaissance supérieurs à ceux obtenus par le modèle de base BM.



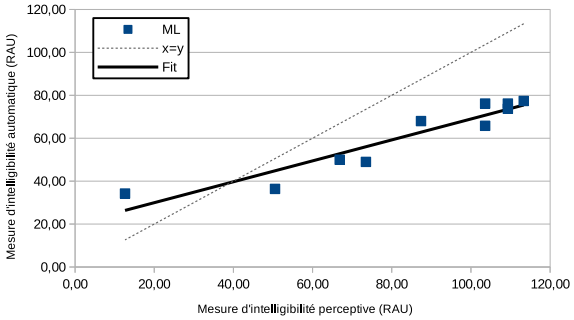

(a) Taux de reconnaissance du mot en première position



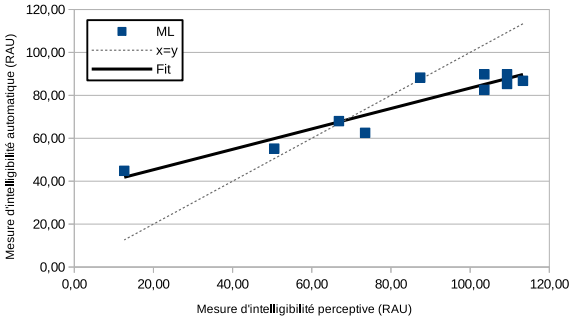

(b) Taux de reconnaissance dans les 10 premières positions

FIGURE 3: Taux de reconnaissances automatique (en utilisant deux modèles de langage) et perceptive de différents niveaux simulés de surdité professionnelle.

En utilisant les deux stratégies de reconnaissance automatique, le comportement de l'approche automatique suit la même tendance que l'évaluation perceptive : plus le niveau de dégradation simulée est élevée, plus le système de RAP a du mal à reconnaître les mots prononcés. En observant de plus prés l'évolution des taux de reconnaissance selon la sévérité de la dégradation simulée, nous retrouvons que la perception humaine arrive à compenser les distorsions observées sur les 5 premiers niveaux de dégradation. À partir du niveau 5, l'intelligibilité perceptive subit une perte, presque linéaire, avec une importante chute de taux de reconnaissance entre les niveaux 9 (50,5%) et 10 (14%). La mesure d'intelligibilité automatique présente une évolution différente, plus sensible aux faibles dégradations (diminution du taux reconnaissance entre les niveaux 4 et 5) et plus résistante au dégradation sévère du niveau 10.




(a) Taux de reconnaissance du mot en première position.




(b) Taux de reconnaissance dans les 10 premières positions.

FIGURE 4: Distribution de la mesure d'intelligibilité automatique du modèle de langage ML « transformé » (RAU) en fonction de la mesure d'intelligibilité perceptive « transformée » (RAU).

Afin de mesurer la capacité de la mesure d'intelligibilité automatique à produire des mesures comparables à celles d'une évaluation perceptive, nous avons calculé le taux de corrélation de Pearson entre les deux mesures. En vue d'éviter les effets plafonds des scores, une transformation en des unités arc-sinus rationnels (rational-arcsine unit - RAU) (Studebaker, 1985) a été réalisée. La figure 4

montre la relation entre les scores perceptifs et automatiques ainsi transformés. Le tableau 2 rapporte les taux de corrélation entre les mesures d'intelligibilité automatique et perceptive.

TABLE 2: Coefficients de corrélation de Pearson observés en comparant les mesures d'intelligibilité automatique et perceptive (p<0,001).

| | Stratégies de comparaison | |
|---|---|---|
| Modèle | Mot reconnu en première position | Mot reconnu dans les 10 premières positions |
| BM | 0,91 | 0,90 |
| ML | 0,94 | 0,94 |

Les taux de corrélation entre les deux mesures d'intelligibilité atteignent 0,94 dans le cas de l'utilisation du modèle de langage ML. Cette mesure montre la capacité de l'approche automatique à répliquer le comportement de la perception humaine face à de la parole simulant différents niveaux de surdité professionnelle. De plus, ce taux est comparable à celui obtenue lors de notre étude précédente sur la presbyacousie : corrélation de 0,97 obtenue dans (Fontan *et al.*, 2017).

Ce lien entre les mesures d'intelligibilité perceptive et automatique confirme le potentiel de l'approche automatique proposée et son potentiel d'utilisation lors des phases de réglage et d'adaptation des audioprothèses destinées aux patients souffrant de la surdité professionnelle.

# 5 Conclusions

Ce travail vise à étudier et analyser les effets de la surdité professionnelle sur l'intelligibilité de la parole. Une évaluation perceptive a démontré que ces effets sont indépendants du genre du locuteur (homme ou femme) ainsi que de son âge (adulte ou enfant). Cette évaluation a aussi permis d'établir une référence à laquelle comparer la mesure d'intelligibilité automatique. Le comportement de cette dernière, basée sur un système de RAP, s'est révélé similaire et comparable à la perception humaine (corrélation de Pearson de 0,94 entre les deux mesures). Ce résultat valide l'intérêt de la méthodologie proposée et son potentiel d'utilisation dans le cadre du réglage et de l'adaptation des audioprothèses aux besoins spécifiques de chaque utilisateur.

De futurs travaux permettrons de généraliser cette approche à d'autre langue (Anglais) et d'étudier sa robustesse face à des conditions de dégradation plus difficile (parole en milieu bruité). Finalement, et vue le lien entre la perception de la parole et sa production, il nous semble intéressant d'exploiter cette mesure d'intelligibilité automatique dans le cadre d'évaluation des troubles de production de la parole tels que la dysarthrie.

# Remerciements

Cette étude a été réalisée dans le cadre du projet numéro 14054092 « CLE 2015 PHONICS : Intelligent Electronic Device for Measuring Speech Comprehension », financé par la région Midi-Pyrénées. Le projet est porté par la Maison des Sciences de l'Homme et de la Société de Toulouse (MSHS-T) en partenariat avec l'IRIT et la société Archean Technologies.

# Références


DELÉGLISE P., ESTEVE Y., MEIGNIER S. & MERLIN T. (2005). The LIUM speech transcription system : a CMU Sphinx III-based system for french broadcast news. In *Ninth European Conference on Speech Communication and Technology*.

FONTAN L., FERRANÉ I., FARINAS J., PINQUIER J., TARDIEU J., MAGNEN C., GAILLARD P., AUMONT X. & FÜLLGRABE C. (2017). Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss. *Journal of Speech, Language, and Hearing Research*, **60**(9), 2394–2405.

FOURNIER J. E. (1951). Audiométrie vocale : les épreuves d'intelligibilité et leurs applications au diagnostic, à l'expertise et à la correction prothétique des surdités.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ESTER 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Tenth Annual Conference of the International Speech Communication Association*.

GOPINATH B., WANG J. J., SCHNEIDER J., BURLUTSKY G., SNOWDON J., MCMAHON C. M., LEEDER S. R. & MITCHELL P. (2009). Depressive symptoms in older adults with hearing impairments : the blue mountains study. *Journal of the American Geriatrics Society*, **57**(7), 1306–1308.

HERMANSKY H. (1990). Perceptual linear predictive (PLP) analysis of speech. *the Journal of the Acoustical Society of America*, **87**(4), 1738–1752.

HUSTAD K. C. & CAHILL M. A. (2003). Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology*, **12**(2), 198–208.

MCBRIDE D. & WILLIAMS S. (2001). Audiometric notch as a sign of noise induced hearing loss. *Occupational and Environmental Medicine*, **58**(1), 46–51.

NEJIME Y. & MOORE B. C. (1997). Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise. *The Journal of the Acoustical Society of America*.

NEW B., BRYSBAERT M., VERONIS J. & PALLIER C. (2007). The use of film subtitles to estimate word frequencies. *Applied psycholinguistics*, **28**(4), 661–677.

SEYMORE K., CHEN S., DOH S., ESKENAZI M., GOUVEA E., RAJ B., RAVISHANKAR M., ROSENFELD R., SIEGLER M., STERN R. *et al.* (1998). The 1997 CMU Sphinx-3 English broadcast news transcription system. In *DARPA Broadcast News Transcription and Understanding Workshop*.

STRAWBRIDGE W. J., WALLHAGEN M. I., SHEMA S. J. & KAPLAN G. A. (2000). Negative consequences of hearing impairment in old age : a longitudinal analysis. *The Gerontologist*, **40**(3), 320–326.

STUDEBAKER G. A. (1985). A rationalized arcsine transform. *Journal of Speech, Language, and Hearing Research*, **28**(3), 455–462.

VESTERGAARD KNUDSEN L., ÖBERG M., NIELSEN C., NAYLOR G. & KRAMER S. E. (2010). Factors influencing help seeking, hearing aid uptake, hearing aid use and satisfaction with hearing aids : A review of the literature. *Trends in amplification*, **14**(3), 127–154.

WEGMANN S., MCALLASTER D., ORLOFF J. & PESKIN B. (1996). Speaker normalization on conversational telephone speech. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, p. 339–341 : IEEE.

XXXIIe Journées d'Études sur la Parole
4-8 juin 2018, Aix-en-Provence, France

# Perception des consonnes et voyelles nasales en parole vocodée : Analyse de la contribution des niveaux de résolution spectrale et temporelle.

Olivier Crouzet[1,2]
(1) LLING – Laboratoire de Linguistique de Nantes – UMR6310 Université de Nantes / CNRS, chemin de la Censive du Tertre, 44312 Nantes Cedex, France
(2) ENT Department – University Medical Center Groningen, Rijksuniversiteit Groningen, Pays-Bas
olivier.crouzet@univ-nantes.fr

## Résumé

Nous présentons une série d'expériences dans lesquelles nous étudions l'impact des niveaux de résolution spectrale et temporelle de signaux de parole vocodée par canaux de bruit sur la classification des consonnes et voyelles du français en nous intéressant plus particulièrement à la comparaison orales / nasales. Les réponses perceptives ont été recueillies séparément pour l'identification des consonnes et des voyelles dans des tâches de classification à choix forcé à N alternatives. Nous étudions dans un premier temps la relation entre les performances de classification et les niveaux de résolution spectrale et temporelle des signaux. Nous présentons ensuite des analyses d'entropie mutuelle qui permettent d'évaluer le degré de préservation des différents « traits » associés aux catégories sonores. Ces résultats font ressortir des difficultés particulières posées par le trait de nasalité pour la classification des voyelles. Un certain nombre de questions méthodologiques et théoriques émergent de ces données et sont discutées.

## Abstract

**Perceptual classification of nasal consonants and vowels in vocoded speech: Contribution of spectral and temporal resolution levels.**

A series of experiments is described in which we investigated the impact of spectral and temporal resolution of channel-vocoded speech on French consonants and vowels, focusing our main interests on a comparison of oral and nasal segments. Participants provided perceptual responses in N-Alternative Forced-Choice tasks focusing on either consonant or vowel identification. We first discuss overall classification performance levels in relation to spectral and temporal resolution of signals. We then turn to mutual entropy analyses which let us estimate information transmission preservation for individual featural categories. From these analyses, it is observed that specific difficulties seem to be associated with nasal vowels. Further methodological and theoretical issues are then discussed.

Mots-clés : parole vocodée ; implants cochléaires ; résolution spectrale ; résolution temporelle ; perception ; théorie de l'information ; entropie mutuelle ; matrices de confusion.

Keywords: vocoded speech; cochlear implants; spectral resolution; temporal resolution; perception; information theory; mutual entropy; confusion matrices.


10.21437/JEP.2018-46

# 1 Introduction

Les types de stimulation produits par les implants cochléaires sont décrits comme transmettant une information de type « variations d'énergie intra-bande » (Shannon *et al.*, 1995) et induisent donc une perte d'information spectrale par rapport aux capacités naturelles du système auditif puisque la résolution spectrale de l'implant dépend essentiellement du nombre d'électrodes implantées (ce nombre étant actuellement lui-même limité par des contraintes techniques liées notamment à la taille de ces électrodes et aux risques de diffusion de potentiel –en anglais « current spread »– liés à cette taille). Néanmoins, on observe depuis les travaux princeps de Van Tasell *et al.* (1987) que des formes acoustiques de ce type (signaux à canaux vocodés) sont en mesure de fournir des informations relativement efficaces pour la classification phonétique. Les données plus récentes portant sur la modélisation du traitement des signaux par les implants cochléaires conduisent à considérer que la « qualité de la résolution spectrale » n'est pas primordiale pour la reconnaissance de la parole (Shannon *et al.*, 1995; Kanedera *et al.*, 1999) et que l'information phonétique serait principalement portée par les « fréquences de modulation d'amplitude » (Kanedera *et al.*, 1999; Christiansen & Greenberg, 2010) qui sont associées aux canaux spectraux, lesquelles correspondent aux modulations d'énergie de basse fréquence qui caractérisent les variations d'amplitude de l'onde dans un canal spectral.

Chez les patients porteurs d'implants cochléaires, outre des variations individuelles importantes en termes de « récupération » rendue possible par l'appareillage, l'une des propriétés qui semble résister à la prise en charge orthophonique est la nasalité. Ces problèmes correspondent aussi bien à des phénomènes de « qualité de la voix » (hyper– / hypo–nasalisation ; Fletcher *et al.*, 1999; Baudonck *et al.*, 2015) que de discrimination phonologique en perception (Borel, 2015, pour les voyelles du français).

Même si les aspects articulatoires de la nasalité peuvent parfois paraître relativement simples, aussi bien la question des propriétés articulatoires des distinctions orales / nasales (Delvaux, 2012; Demolin *et al.*, 2003; Carignan *et al.*, 2015) que celle des effets acoustiques de la nasalité (House & Stevens, 1956; Maeda, 1982; Stevens, 1998; Feng & Castelli, 1996; Rossato, 2000) continuent de poser des problèmes cruciaux en termes de modélisation théorique des relations articulatoire–acoustique et des mécanismes de perception. Or il se trouve que les analyses phonétiques qui sont proposées dans le cadre du modèle source-filtre suggèrent une complexité du spectre de sortie très marquée, laquelle s'explique par le phénomène de couplage entre les cavités orale et nasale (Maeda, 1982; Stevens, 1998, cf. Fig. 1). Si les propriétés des consonnes et des voyelles nasales diffèrent (en raison des contributions relatives différentielles des cavités orale et nasale et de leur organisation temporelle), les mécanismes fondamentaux impliqués peuvent générer certains phénomènes acoustiques similaires comme l'élargissement de la largeur de bande des formants ou l'atténuation de l'énergie du signal. Certains de ces mécanismes pourraient constituer un frein à la transmission des informations acoustiques pertinentes dans un implant cochléaire.

Afin d'évaluer les contributions respectives des paramètres de résolution spectrale et temporelle associés à la perception des nasales, nous avons mis en place une série d'expériences dans lesquelles nous étudions la catégorisation d'un sous ensemble des consonnes et voyelles du français dans deux tâches distinctes (une tâche de classification des consonnes, une de classification des voyelles). Notre objectif est d'évaluer le rôle de ces deux domaines de résolution sensorielle pour la classification des catégories sonores et d'étudier plus particuliè-

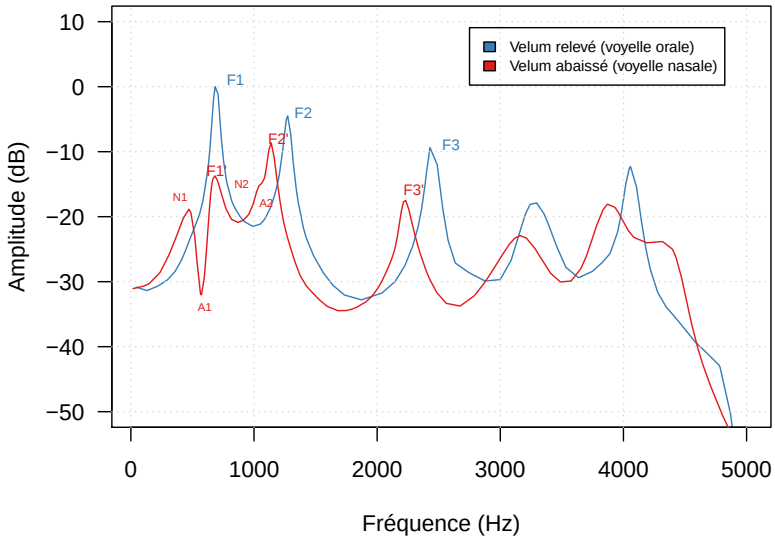



FIGURE 1: Modélisation articulatoire–acoustique de l'abaissement du velum pour la production des voyelles proposée par Maeda (1982). Cette simulation met en évidence l'impact du couplage oral–nasal et des anti-résonances ($A_1$, $A_2$) qui en découlent sur le spectre de sortie : diminution de l'amplitude de certains formants « oraux »($F_1 \rightarrow F_1'$, $F_2 \rightarrow F_2'$, $F_3 \rightarrow F_3'$), décalages en fréquence, combinaison de résonances « orales » et « nasales » ($N_1$, $N_2$) et des anti-résonances provoquant une augmentation des largeurs de bandes. Graphique adapté de Maeda (1982).

rement le comportement des catégories nasales dans les performances de classification. Ces données pourraient contribuer à améliorer le traitement des informations acoustiques pour les personnes déficientes auditives portant un implant cochléaire mais fourniraient aussi des informations essentielles pour la modélisation acoustique des nasales et la compréhension des mécanismes perceptifs qui leur sont associés.

# 2 Méthode

Deux expériences parallèles ont été conçues afin d'évaluer les impacts perceptifs des propriétés de résolution spectrale et temporelle de la parole sur l'identification des consonnes et voyelles nasales. Dans la première expérience on étudie la classification perceptive de consonnes dans un contexte VCV. Dans la seconde expérience, l'identification de voyelles est abordée à travers la classification de segments vocaliques isolés ne présentant pas de transitions formantiques. Dans les deux cas, les participants réalisent une tâche de classification à choix forcé à N alternatives.

## 2.1 Expérience 1 : Consonnes

### 2.1.1 Participants

Dix-neuf participants volontaires adultes n'ayant pas de troubles auditifs connus ont pris part à l'expérience. Les résultats de 3 d'entre eux ont été retirés des données en raison de

problèmes techniques n'ayant pas permis de réaliser l'intégralité de l'expérience. Les résultats des 16 participants restants sont présentés.

### 2.1.2 Matériel

Les stimuli sont des séquences VCV [1] sans signification. La consonne (choisie parmi 19 consonnes du français dans l'ensemble {b, d, g, p, t, k, v, z, ʒ, f, s, ʃ, l, ʁ, w, j, m, n, ɲ}) est produite dans 3 contextes vocaliques différents ({i, a, u}). Les stimuli ont été produits par une locutrice adulte et numérisés sans compression à une fréquence d'échantillonnage de 16 kHz avec un taux de quantification de 16 bits. Ils ont ensuite été traités par un vocodeur à canaux de bruits développé dans l'environnement Octave (Eaton *et al.*, 2015) en faisant varier 2 paramètres : le nombre de canaux spectraux (2, 4, 6, 8 canaux de bruit) et la fréquence de coupure des modulations d'amplitude (filtre passe-bas ; 4, 16, 128 Hz). Dans la condition 4 Hz, seules les fréquences de modulation lentes sont préservées. Pour la fréquence de coupure 16 Hz, les fréquences de modulation lentes et moyennement rapides sont préservées. À 128 Hz, toutes les fréquences de modulation sont préservées. Les fréquences de coupure des modulations d'amplitude ont été choisies sur la base d'évaluations subjectives informelles. Les fréquences du banc de filtres passe-bande utilisé pour créer les canaux de bruit suivent une progression régulière en ERB (Moore & Glasberg, 1983) selon l'implémentation de Slaney (1993).

### 2.1.3 Procédure

Le recueil des données se faisait dans une pièce calme. Les stimuli étaient présentés à travers un casque à un niveau sonore jugé confortable par les participants. L'expérience débutait par une familiarisation avec la tâche. Les stimuli étaient ensuite présentés dans un ordre aléatoire. Un tableau représentant les consonnes sous forme orthographique était affiché sur l'écran de l'ordinateur sur lequel l'expérience était réalisée. Les participants avaient pour tâche de cliquer sur la case qui contenait la consonne identifiée. Il n'était pas possible de réécouter le stimulus. Dès que la réponse était donnée, le stimulus suivant était diffusé. Les participants pouvaient faire autant de pauses qu'ils le souhaitaient pendant la session. Chaque stimulus était présenté 1 fois. Au total, chaque participant donnait 684 réponses (19 consonnes $\times$ 3 voyelles $\times$ 4 conditions de nombre de canaux spectraux $\times$ 3 conditions de résolution temporelle).

### 2.1.4 Résultats

Les taux de reconnaissance correcte ont été calculés sur l'ensemble de l'expérience mais les résultats sont présentés en séparant les consonnes orales et les consonnes nasales de manière à évaluer les différences potentielles entre ces deux catégories. Ces résultats sont représentés dans la figure 2. Nous étudions l'impact de la résolution spectrale et temporelle sur les performances de classification. Un test binomial a été appliqué pour chaque condition afin de déterminer si le taux de réponses correctes dépasse significativement la probabilité aléatoire de réponses correctes ($1/19 \times 100 = 5.26\%$). Les nasales étant moins nombreuses que les orales, le seuil de significativité est plus haut pour les nasales.

1. Voyelle-Consonne-Voyelle

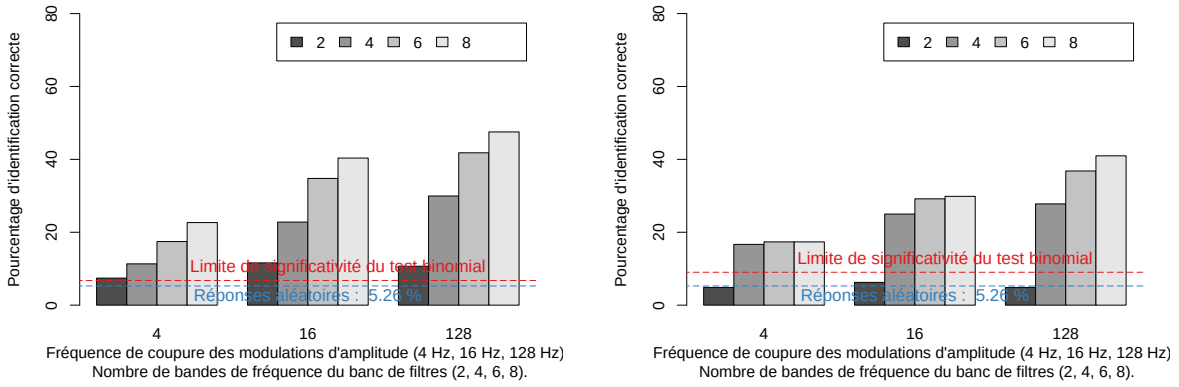



FIGURE 2: Performances de classification des consonnes orales (Fig. 2a) et nasales (Fig. 2b). Le trait horizontal bleu indique le pourcentage de réponses correctes correspondant au taux théorique de réponses au hasard dans la tâche de classification à choix forcé. Le trait rouge caractérise le niveau de pourcentage de réponses correctes au-delà duquel les performances mesurées dépassent significativement (pour un test binomial et un seuil $p < 0.05$) le taux aléatoire de réponses correctes.

Globalement, on observe que la performance progresse conjointement à l'amélioration de la résolution spectrale et temporelle. La plupart des conditions permettent de dépasser le taux de réponses correctes aléatoire. C'est le cas dans toutes les conditions pour les consonnes orales. Pour les consonnes nasales, seule la condition la plus dégradée (2 canaux quelles que soient les fréquences de modulation disponibles) empêche les participants de dépasser le seuil de réponses au hasard. Dans toutes les autres conditions et de manière similaire aux consonnes orales, la performance est significativement supérieure au hasard et s'améliore avec l'accroissement des niveaux de résolution spectrale et temporelle. On voit donc que la difficulté de classification des consonnes nasales est légèrement plus marquée que pour les orales mais que globalement la forme de la progression est similaire. Cette légère différence pourrait notamment s'expliquer par la plus faible proportion de nasales dans la langue et / ou dans le matériel de l'expérience ainsi que par de plus grandes difficultés à traiter l'information acoustique propre aux nasales.

## 2.2 Expérience 2 : Voyelles

### 2.2.1 Participants

Les mêmes dix-neuf participants ont participé à l'expérience au cours de la même session. Les résultats de 2 d'entre eux ont été retirés des données en raison de problèmes techniques et seules les données des 17 participants restants sont présentées.

### 2.2.2 Matériel

Les stimuli sont des voyelles isolées qui ont été produites par la même locutrice au cours de la même session. Lors de l'enregistrement, des mots contenant ces voyelles dans leur

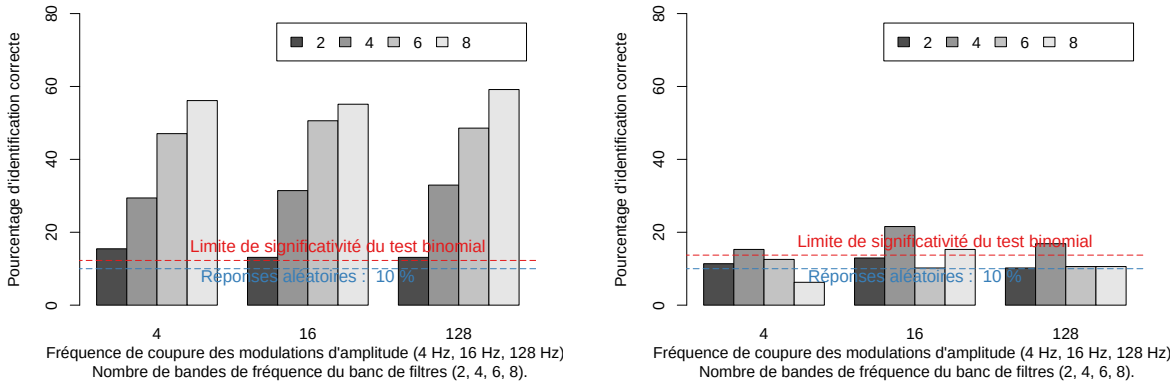



FIGURE 3: Performances de classification des voyelles orales (Fig. 3a) et nasales (Fig. 3b).

première syllabe (ouverte) étaient présentés à l'écran et la locutrice devait produire le mot en maintenant la réalisation de la première voyelle pendant plusieurs secondes. Les enregistrements ont ensuite été édités pour extraire une portion spectralement stable de 250 ms de chaque voyelle en appliquant une enveloppe d'accroissement puis d'atténuation progressive et rapide (10ms de temps de montée / descente) de l'amplitude au début et à la fin du signal. Les 10 voyelles utilisées sont : {i, e, a, y, ø, u, o, ɛ̃, ɑ̃, ɔ̃}.

Les stimuli ont été traités selon les mêmes principes que dans l'expérience 1. Il est important de noter que pour cette expérience, il n'est attendu aucun impact des fréquences de coupure des modulations d'amplitude puisqu'il n'y a normalement aucune information temporelle distinctive dans ces voyelles : elles sont stables du point de vue du spectre et de l'énergie. Nous avons choisi de conserver les mêmes conditions dans la perspective d'une étude ultérieure des propriétés dynamiques des voyelles.

### 2.2.3 Procédure

La procédure était la même que dans l'expérience 1. L'interface graphique utilisée affichait des équivalents orthographiques des voyelles pour recueillir les réponses des participants. Afin d'obtenir approximativement la même quantité de données dans les deux expériences, chaque stimulus était répété 5 fois au cours de l'expérience. Au total, chaque participant donnait 600 réponses (10 voyelles $\times$ 4 conditions de nombre de canaux spectraux $\times$ 3 conditions de résolution temporelle $\times$ 5 répétitions).

### 2.2.4 Résultats

Nous avons procédé aux mêmes analyses que dans l'expérience 1. Les résultats observés respectivement pour les voyelles orales et nasales sont présentés dans la figure 3. Le taux théorique de réponses au hasard est de $1/10 \times 100 = 10\%$. De même qu'avec les consonnes, les voyelles nasales étant moins nombreuses que les voyelles orales, le seuil de significativité est plus haut pour les nasales

Pour les voyelles orales, on observe un comportement des réponses en fonction des niveaux de résolution spectrale et temporelle tout à fait similaire aux consonnes orales. À l'inverse, les performances observées pour les voyelles nasales sont fortement dégradées. D'une part, très peu de conditions donnent lieu à des taux de reconnaissance correcte qui dépassent significativement le seuil de réponses au hasard. Seules la condition à 4 canaux (pour toutes les fréquences de modulation d'amplitude) et la condition à 8 canaux pour la fréquence de coupure 16 Hz donnent lieu à des performances dépassant le seuil de significativité du test binomial. D'un point de vue global, l'amélioration progressive des performances qui est observée sur les consonnes (orales et nasales) et sur les voyelles orales avec l'accroissement de la résolution spectrale est totalement absente des résultats observés sur les voyelles nasales.

### 2.3 Analyses d'entropie mutuelle

Afin d'affiner l'interprétation des données de performance, nous avons procédé à l'analyse quantitative des matrices de confusion à travers des calculs d'entropie mutuelle (qu'on appelle aussi « Taux de transfert d'information » à travers un canal / un medium ; Shannon, 1948; Miller & Nicely, 1955; Christiansen & Greenberg, 2012) en nous focalisant sur des traits articulatoires permettant de regrouper les segments en catégories. Afin de compenser les problèmes d'interprétation liés aux déséquilibres des effectifs des différentes catégories et au nombres de catégories distinctes, nous avons calculé l'entropie mutuelle *relative* (ou « taux de transfert d'information normalisé »).

Par manque de place, il n'est pas possible de détailler ces résultats ici. Les mesures effectuées confirment la difficulté à différencier les voyelles orales des nasales dans toutes les conditions de résolution étudiées. La différenciation entre consonnes orales et nasales, correcte mais néanmoins assez limitée, semble comparable à la classification de la place d'articulation ou du mode, le voisement étant l'information la mieux perçue. Pour les voyelles, les propriétés d'aperture, de position et d'arrondissement semblent aussi bien transmises que le voisement pour les consonnes.

Globalement, il semble donc que l'information de nasalité des voyelles soit très nettement dégradée en parole vocodée et ce, quelles que soient les conditions de résolution spectrale étudiées dans l'expérience.

## 3 Discussion

Les résultats de performance semblent mettre en évidence une forte difficulté des participants à identifier correctement les voyelles nasales. Cette observation est notamment soulignée par la forme globale des résultats : il ne semble y avoir aucune amélioration progressive de la performance de classification avec le niveau de résolution spectrale. Le fait que certaines conditions donnent lieu à des taux de réponse qui dépassent significativement le seuil de réponses au hasard pourrait être la conséquence de résultats statistiques aléatoires liés au risque d'erreur de Type I. Néanmoins, il est intéressant de constater que cette situation s'observe de manière cohérente pour les 3 fréquences de coupure des modulations d'amplitude affectées aux stimuli composés de 4 canaux. Or ce chiffre pourrait correspondre au nombre de canaux idéalement efficaces pour représenter les régularités statistiques de la parole (Ming &

Holt, 2009). Les fréquences limites des filtres utilisés dans le vocodage des signaux à 4 canaux (autour de 600, 1400, 4000 Hz pour les 3 frontières intermédiaires) semblent assez proches des limites identifiées par Ueda & Nakajima (2017). À l'issue d'une analyse factorielle réalisée sur des signaux de parole issus de 8 langues différentes, Ueda & Nakajima (2017) déduisent 4 « bandes spectrales idéales » qui seraient *optimales* pour porter l'information acoustique et les 3 frontières identifiées par les auteurs sont respectivement 540, 1720 et 3300 Hz. Cette hypothèse devra être approfondie car elle pourrait fournir une piste essentielle pour l'étude de la transmission des informations nasales dans un implant.

Les résultats des analyses d'entropie mutuelle qui n'ont été que partiellement évoqués ici semblent plutôt suggérer que l'information de distinction orale / nasale associée aux voyelles est très fortement dégradée pour l'ensemble des conditions de résolution spectrale et temporelle. Il reste cependant que le *design* mis en œuvre ne permet pas de comparer à la fois l'analyse perceptive de la distinction orale / nasale et celle de la différenciation des nasales entre elles. En effet, les mesures d'entropie mutuelle associées à des paramètres comme l'arrondissement, l'aperture ou la position pour les voyelles intègrent nécessairement les résultats des voyelles orales. Nous prévoyons d'étudier spécifiquement le comportement des voyelles orales et nasales dans un cadre qui permettra d'obtenir des données pertinentes pour étudier cette question en mettant en place un design spécifiquement adapté. Par ailleurs, il semble essentiel de généraliser les résultats de cette étude à d'autres locuteurs et à des conditions de résolution spectrale plus fines.

# Remerciements

Ce travail a reçu le soutien financier du Conseil Scientifique de l'Université de Nantes (Programme « Interdisciplinarités »).

# Références

Baudonck N., Van Lierde K., D'Haeseleer E. & Dhooge I. (2015). Nasalance and nasality in children with cochlear implants and children with hearing aids. *International Journal of Pediatric Otorhinolaryngology*, **79**(4), 541–545.

Borel S. (2015). *Perception auditive, visuelle et audiovisuelle des voyelles nasales par les adultes devenus sourds. Lecture labiale, implant cochléaire, implant du tronc cérébral.* PhD thesis, Université de la Sorbonne Nouvelle – Paris 3.

Carignan C., Shosted R. K., Fu M., Liang Z.-P. & Sutton B. P. (2015). A real-time MRI investigation of the role of lingual and pharyngeal articulation in the production of the nasal vowel system of French. *Journal of Phonetics*, **50**, 34–51.

Christiansen T. U. & Greenberg S. (2010). Frequency selective filtering of the modulation spectrum and its impact on consonant identification. In *Linguistic Theory and Raw Sound*, volume 40 of *Copenhagen Studies in Language*, p. 119. Copenhagen, DK : Samfundslitteratur.

Christiansen T. U. & Greenberg S. (2012). Perceptual Confusions Among Consonants, Revisited : Cross-Spectral Integration of Phonetic-Feature Information and Consonant

Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, **20**(1), 147–161.

Delvaux V. (2012). *Les voyelles nasales du français : Aérodynamique, articulation, acoustique et perception*. Peter Lang, Éditions Scientifiques Internationales.

Demolin D., Delvaux V., Metens T. & Soquet A. (2003). Determination of velum opening for french nasal vowels by magnetic resonance imaging. *Journal of Voice*, **17**(4), 454–467.

Eaton J. W., Bateman D., Hauberg S. & Wehbring R. (2015). *GNU Octave version 4.0.0 manual : a high-level interactive language for numerical computations.* UK : Network Theory Limited. 2002.

Feng G. & Castelli E. (1996). Some acoustic features of nasal and nasalized vowels : A target for vowel nasalization. *The Journal of the Acoustical Society of America*, **99**(6), 3694–3706.

Fletcher S., Mahfuzh F. & Hendarmin H. (1999). Nasalence in the speech of children with normal hearing and children with hearing loss. *American Journal of Speech Language Pathology*, **8**, 241–248.

House A. S. & Stevens K. N. (1956). Analog studies of the nasalization of vowels. *Journal of Speech and Hearing Disorders*, **21**(2), 218–232.

Kanedera N., Arai T., Hermansky H. & Pavel M. (1999). On the relative importance of various components of the modulation spectrum for automatic speech recognition. *Speech Communication*, **28**, 43–55.

Maeda S. (1982). The role of the sinus cavities in the production of nasal vowels. In *ICASSP – IEEE International Conference on Acoustics Speech and Signal Processing*, volume 7, p. 911–914.

Miller G. A. & Nicely P. E. (1955). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, **27**(2), 338–352.

Ming V. L. & Holt L. L. (2009). Efficient coding in human auditory perception. *The Journal of the Acoustical Society of America*, **126**(3), 1312–1320.

Moore B. C. J. & Glasberg B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*, **74**, 750–753.

Rossato S. (2000). *Du son au geste, inversion de la parole : le cas des voyelles nasales.* Thèse de doctorat, Université Sthendal, Grenoble, France.

Shannon C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 399–423, 623–656.

Shannon R., Zeng F., Kamath V., Wygonski J. & Ekelid M. (1995). Speech recognition with primarily temporal cues. *Science*, **270**, 303–304.

Slaney M. (1993). *An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank.* Rapport interne 35, Apple Computer.

Stevens K. N. (1998). *Acoustic Phonetics.* Cambridge, Mass., USA : The MIT Press.

Ueda K. & Nakajima Y. (2017). An acoustic key to eight languages/dialects : Factor analyses of critical-band-filtered speech. *Scientific Reports*, **7**, 42468.

Van Tasell D., Soli S., Kirby V. & Widin G. (1987). Speech waveform envelope cues for consonant recognition. *The Journal of the Acoustical Society of America*, **77**, 1069–1077.

# Réduction de la coarticulation et vieillissement

Daria D'Alessandro & Cécile Fougeron
Laboratoire de Phonétique et Phonologie - UMR 7018
Université Sorbonne Nouvelle - CNRS
19 Rue des Bernardins 75005 PARIS - France
daria.dalessandro@etud.sorbonne-nouvelle.fr,cecile.fougeron@univ-paris3.fr

## RESUME

La parole évolue au cours de la vie avec le développement, chez les enfants, mais aussi avec le vieillissement, chez les adultes. Outre des modifications temporelles et spectrales co nnues, cette étude vise à examiner les effets du vieillissement sur la coarticulation anticipatoire voyelle_à_voyelle en français. Cet effet est testé en fonction de la durée des voyelles et de différences entre variantes régionales. Les productions de 167 locuteurs couvrant trois groupes d'âge (20-39) (50-69) (70-89) et 4 variantes régionales sont examinées. La coarticulation de V2 (/a/ ou /i/) sur V1 (/a/) est mesurée en termes d'abaissement de $F_1$ et montée de $F_2$. Les résultats montrent un degré de coarticulation et un allongement des voyelles variables en fonction de la variante régionale ainsi qu'un allongement global des voyelles chez les locuteurs les plus âgés. Plus intéressant, ces variations s'accompagnent d'une diminution de la coarticulation avec l'âge.

## ABSTRACT

**Reduction in coarticulation and aging**
In the course of life, speech varies with respect to children's development but also with respect to aging. Besides the temporal and spectral modifications already attested, this study aims to investigate the effect of aging on vowel to vowel anticipatory coarticulation in French. This effect is tested according to vowel duration and to differences in regional varieties. Data from 167 speakers distributed across three age groups (20-39) (50-69) (70-79) and four regional varieties are investigated. The influence of V2 (/a/ or /i/) on V1 (/a/) is measured as a lowering of F1 and a rise of F2. Results show that coarticulation and vowel duration varies with regional variety and that vowels of older speakers are lengthened. More interestingly, results show a reduction of coarticulation with age.

MOTS-CLES : Coarticulation V à V ; vieillissement ; variété régionale du français
KEYWORDS: V to V coarticulation; aging; French regional varieties

# 1 Introduction

La coarticulation peut être définie comme le processus par lequel, dans la parole, un segment est modifié par les segments qui le précèdent (coarticulation persévérante) et/ou qui le suivent (coarticulation anticipatoire). Cette influence mutuelle entre les segments ne s'exerce pas seulement entre segments adjacents mais s'observe aussi à distance. La coarticulation a été beaucoup étudiée car


10.21437/JEP.2018-47

elle soulève des questions relatives à l'organisation et la planification de la parole. En effet, si la coarticulation intra-syllabique et la coarticulation persévérante ont été attribuées, au moins en partie, à des contraintes biomécaniques, la coarticulation anticipatoire à distance a été vue comme la résultante de phénomènes anticipatoires lors de la planification de la parole (Whalen, 1990).

Un cas de coarticulation à distance est la coarticulation voyelle à voyelle (V-à-V) : une voyelle exerce une influence sur une voyelle qui la suit ou qui la précède au travers des consonnes qui les séparent. Cette coarticulation V-à-V dépend de différents facteurs, comme la nature de la (ou des) consonne entre la source et la cible. Ainsi, le degré de coarticulation V-à-V serait inversement proportionnel au degré de contact linguopalatal et d'implication du dos de la langue dans la séquence (Recasens, 1989 ; Mok, 2001). Une consonne palatale ou vélaire bloquerait d'avantage la coarticulation qu'une consonne bilabiale ; pareillement, les voyelles fermées /i/ et /u/ seraient plus résistantes à la coarticulation que les voyelles ouvertes. Dans notre étude nous allons donc exploiter ces caractéristiques en nous focalisant sur des cas favorables à la coarticulation anticipatoire de $V_2$ sur $V_1$ dans des séquences /ap$V_2$/.

Un facteur qui peut affecter la coarticulation est aussi l'âge du locuteur. L'effet de l'âge sur la coarticulation dans la parole a été, à notre connaissance, exclusivement focalisé sur la comparaison entre des groupes d'enfants et d'adultes d'âge moyen (et, dans une moindre mesure, entre adolescents et adultes), dans le cadre du développement de la parole. Les differentes études sur le sujet montrent globalement un changement dans les patrons de coarticulation entre enfants et adultes et une variabilité de la coarticulation plus forte chez les enfants par rapport aux adultes (*inter alia* Zharkova, 2012 ; Noiray, Ménard et Iskarous, 2013 ; Barbier *et al.*, 2015). Toutefois, la direction du changement de la coarticulation n'est pas claire : certains montrent une coarticulation majorée chez les enfants par rapport aux adultes (Nittrouer *et al.*, 1996 ; Zharkova *et al*,. 2011), d'autres montrent une coarticulation similaire (Rubertus *et al.*, 2016) ou réduite (Zharkova, 2012 ; Barbier *et al.*, 2016).

Si un effet de l'âge sur la coarticulation a été attesté en relation avec le développement du langage chez les enfants, les effets du vieillissement sur la coarticulation sont méconnus. Des effets du vieillissement sur la parole en général ont été documentés dans la littérature. Par exemple, il a été montré chez les personnes âgés un ralentissement de la parole (Jacewicz *et al.*, 2009 ; Bilodeau-Mercure et Tremblay, 2016), une augmentation de la variabilité de la durée des segments ou du VOT (Morris et Brown, 1984), des modifications de la $F_0$ (diminution chez les femmes, augmentation chez les hommes, Torre et Barlow, 2009), une diminution de $F_1$, voire $F_2$ (Xue et Hao, 2003 ; Harrington *et al*., 2007). Ces effets du vieillissement sur la parole peuvent être liés à des changements dans les mouvements en général qui ont été largement documentés. En effet, les effets du vieillissement sur la cinématique des mouvements comprennent un allongement global de leur durée (de 30% à 60% en fonction des tâches motrices), une plus grande variabilité dans les trajectoires et les positions finales des mouvements, et une moins bonne coordination des mouvements (pour une revue, voir Ketcham et Stelmach, 2004). Par exemple Brown (1996) montre chez des adultes âgés (70-95) que la durée de simples mouvements du bras est plus variable, avec une asymétrie entre accélération et décélération du mouvement. Contreras-Vidal et *al.* (1998) observent une diminution de la coordination spatiale et une augmentation de la variabilité temporelle des mouvements avec le vieillissement. On peut donc se poser la question de savoir si ces changements peuvent affecter les temps et la coordination des gestes articulatoires dans la parole.

Connaitre l'évolution de la parole chez l'adulte est essentiel pour la recherche clinique, où il faut normaliser les données selon l'âge, mais aussi pour une meilleure compréhension générale du système de production de la parole. En effet, les changements liés à l'âge peuvent provenir soit de changements

anatomiques et physiologiques dans l'appareil phonatoire, soit de changements neurologiques qui affectent le control moteur ou les fonctions cognitives (Torre et Barlow, 2009 ; Bilodeau-Mercure *et al.,* 2016). C'est pourquoi dans cette étude, nous visons à examiner les effets possibles de l'âge des locuteurs sur la coarticulation anticipatoire V-à-V en français, en fonction aussi de l'origine du locuteur. En outre, nous allons examiner aussi les variations possibles liées à la dimension temporelle de la parole (ici des durées vocaliques) chez la personne âgée et la variabilité temporelle en fonction de la variété régionale, celle-ci pouvant aussi avoir un effet sur l'organisation temporelle de la parole (Jacevicz et al 2010, Verhoeven et al. 2004). Ces premières analyses des effets du vieillissement sur la coarticulation sont conçues comme préliminaires à des analyses sur un plus grand nombre de locuteurs.

# 2 Méthode

Un ensemble de 167 locuteurs, répartis sur 3 classes d'âges, 20-39, 50-69 et 70-89 ans, a été sélectionné dans la base de données MonPaGe_HA (Fougeron *et al.* 2018). Les caractéristiques des groupes d'âge sont présentées dans le tableau I. Les locuteurs (hommes et femmes) sont originaires de 4 régions francophones : région parisienne pour les locuteurs codés FR, région de Mons (Belgique) pour le code BE, région de Genève (Suisse) pour le code CH, région de Montréal (Québec) pour le code QC. Ces locuteurs ont été enregistrés chez eux, par des proches ou des étudiants, selon un protocole dont la passation est informatisée et codifiée. Ce protocole d'évaluation de la parole, MonPaGe, cible diverses dimensions de parole pouvant être altérées en pathologie. Pour cette étude, la partie du protocole construite pour évaluer la production continue lors de la lecture d'un texte (188 mot, env. 1 min.) a été utilisée. Dans ce texte, 7 occurrences du mot 'Papa' et 5 occurrences du mot 'Papi' nous ont permis d'évaluer la coarticulation V-à-V entre $V_1$ (/a/) et $V_2$ (/a/ ou /i/) dans des séquences /p$V_1$p$V_2$/.

Après segmentation manuelle, la durée des voyelles $V_1$ a été mesurée, ainsi que la fréquence de ses F1 et F2. Comme l'effet coarticulatoire est légèrement plus fort sur la deuxième portion de la voyelle une moyenne des mesures à 60%, 70% et 80% de la durée de la voyelle pour F1 et F2 a été retenue. Une mesure composite de la compacité entre F1 et F2, représentée par la différence F2-F1 (toujours prise sur la partie 60-80% de la voyelle) a également été analysée mais ne sera pas présentée ici faute de place.

L'effet du type de $V_2$ (/a/ ou /i/) dont la significativité fera preuve de présence d'une influence coarticulatoire de $V_2$ sur $V_1$ ainsi que les interactions de ce facteur V2 avec l'âge des locuteurs (20-39, 50-69, 70-89), leur origine (BE, CH, FR, QC), et leur sexe (H, F) sont testés à l'aide d'un modèle linéaire mixte construit à l'aide du package de R lme4 (Bates *et al.*, 2015 ; R Core Team, 2017). Les variables dépendantes sont la durée et les F1 et F2 de $V_1$. Les variations liées aux différences inter-locuteurs sont modélisées par une structure aléatoire incluant des interceptes et pentes aléatoires par $V_2$ par locuteur pour les formants, et un intercepte aléatoire pour la variable durée (la pente aléatoire n'améliorant pas le modèle dans ce cas). L'effet des facteurs fixes et de leur interaction est testé par comparaison de modèles avec la fonction *anova.* Les analyses des contrastes à postériori ont été effectuées avec la fonction *lsmeans* (bibliothèque 'emmeans', Lenth *et al.*, 2018).

Table 1 : Effectif (N), Age moyen (en gras), écart type (entre parenthèse), âge minimum et maximum des locuteurs par groupes d'âge, sexe et variété régionale.

| Groupe | Sexe | BE | CH | FR | QC |
|---|---|---|---|---|---|
| Gr1(20-39) | F | N=8; **31** (5), 23<>36 | N=6; **27** (4), 22<>35 | N=8; **32** (6), 23<>37 | N=6; **26** (4), 22<>33 |
| | H | N=8; **30** (4), 25<>36 | N=6; **27** (5), 22<>36 | N=8; **29** (4), 24<>36 | N=6; **28** (5), 22<>35 |
| Gr2(50-69) | F | N=8; **59** (6), 50<>66 | N=6; **59** (7), 51<>69 | N=8; **59** (6), 51<>67 | N=6; **58** (7), 50<>66 |
| | H | N=8; **59** (6), 51<>67 | N=5; **59** (7), 51<>66 | N=8; **58** (7), 50<>68 | N=6; **59** (6), 50<>67 |
| Gr3(70-89) | F | N=8; **79** (7), 70<>89 | N=6; **79** (5), 73<>86 | N=8; **80** (4), 73<>85 | N=6; **78** (7), 70<>86 |
| | H | N=8; **80** (4), 75<>85 | N=6; **79** (8), 70<>88 | N=8; **80** (5), 72<>88 | N=6; **78** (4), 71<>82 |

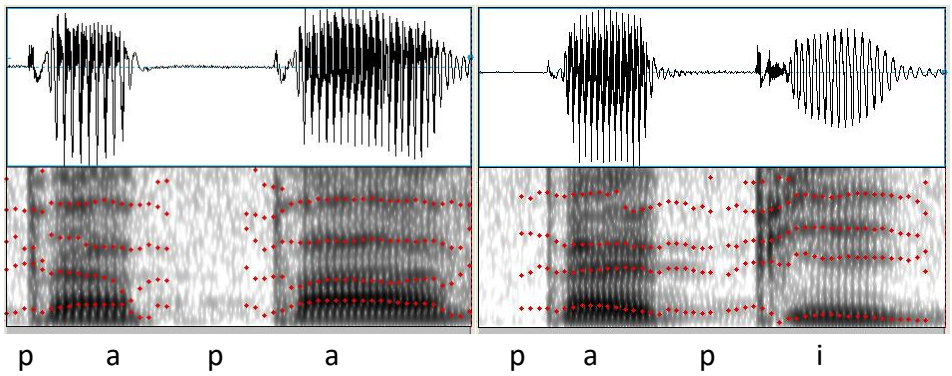



Figure 1 : illustration spectrographique de la variation de $V_1$/a/ en fonction de la voyelle suivante : /a/ dans 'papa' à gauche, /i/ dans 'papi' à droite, produits par la locutrice FR_F_AB9, Gr1.

Table II. Résumé des effets fixes et interactions

| | F1 | F2 | Durée |
|---|---|---|---|
| V2 | $\chi^2(1)=222,7$ ; p<.0001 | $\chi^2(1)=222,7$; p<.0001 | $\chi^2(1)=32,9$ ; p<.0001 |
| AGE | ns | $\chi^2(2)=15,9$; p<.0001 | $\chi^2(2)=45,9$; p<.0001 |
| ORIGINE | ns | $\chi^2(3)=61,8$; p<.0001 | $\chi^2(3)=25,4$ ; p<.0001 |
| SEXE | $\chi^2(1)=100,5$ ; p<.0001 | $\chi^2(1)=98,2$; p<.0001 | ns |
| V2 :AGE :ORIGINE | $\chi^2(12)=34,3$ ; p=.0006 | $\chi^2(12)=31,1$; p=.0002 | ns |
| V2 :AGE :SEXE | $\chi^2(4)=10,1$ ; p=.04 | ns | ns |
| V2 : AGE | $\chi^2(2)=16,1$ ; p=.0003 | $\chi^2(2)=13,4$; p=.001 | ns |
| V2 : ORIGINE | $\chi^2(3)=49,3$ ; p<.0001 | $\chi^2(3)=63,8$; p<.0001 | $\chi^2(3)=13,4$, p<.0001 |
| V2 : SEXE | $\chi^2(1)=6,6$ ; p=.009 | ns | ns |

# 3 Résultats

L'effet contextuel de $V_2$/i/ sur $V_1$/a/ se traduit par un abaissement de F1 et une élévation de F2 par rapport à la condition où $V_1$ est suivi d'un $V_2$/a/, comme l'illustrent les exemples présentés Figure 1. Combinés ces effets ont pour conséquence une diminution de la compacité F2-F1 en contexte V2 /i/. Les effets principaux obtenus sur la mesure de F1 et F2 de $V_1$ sont donnés dans la Table II et illustrés dans la Figure 2. Un effet significatif sur F1 est trouvé pour le facteur $V_2$ avec une baisse significative de F1 en contexte $V_2$(i) (t(167.8)= 11.7 ; p<.0001), et pour le facteur SEXE avec sans surprise des F1

globalement plus bas chez les hommes (t(174.3)= 11.5, p<.0001). Plus intéressant, nous observons une interaction significative entre l'effet de $V_2$ et tous les autres facteurs, et même une interaction triple. Ainsi, l'effet de $V_2$ dépend du groupe d'âge et de l'origine des locuteurs. Pour les locuteurs BE, il n'y a pas d'effet de $V_2$ sur F1 (i.e. pas d'effet contextuel), ceci dans aucun des groupes d'âge. Par contre dans les trois autres variétés régionales, la baisse de F1 dans la condition $V_2$/i/ est significative, mais elle dépend des groupes d'âge : elle est significative pour les 3 groupes d'âge chez les QC (Gr1(t(182.6= 5.4 p <.0001), Gr2 (t(178.6)= 6.5 p <.0001), Gr3 (t(208.9)= 5.2 p<.0001) et les FR (Gr1(t(176.3)= 7,9 p <.0001), Gr2 (t(183,7)= 7,1 p <.0001), Gr3 (t(185,9)= 5.6 p<.0001), mais la différence entre $V_2$/a/ et $V_2$/i/ diminue pour le groupe le plus âgé. Ceci apparaît clairement sur la Figure 2, pour les femmes FR et QC, mais pas pour les hommes. Pour les locuteurs CH, par contre, la baisse de F1 en $V_2$/i/ n'est significative que pour les locuteurs des deux premiers groupes (Gr1(t(176.2)= 8,8 p <.0001), Gr2 (t(180,9)= 3,4 p <.009), alors que le groupe 3 (les plus âgés) se distingue par une absence d'effet contextuel, ce qui est clairement visible sur la Figure 2 pour les deux sexes.

Les effets de $V_2$ sur le F2 de $V_1$ sont illustrés sur la partie droite de la Figure 2. Un effet significatif est trouvé pour tous les facteurs fixes. En effet si le F2 de /a/ est globalement élevé par $V_2$/i/, il varie aussi selon l'âge, l'origine et le sexe du locuteur. Pour ce qui nous intéresse, nous observons que l'effet de $V_2$ interagit avec l'âge et l'origine du locuteur : une augmentation significative de F2 en contexte $V_2$/i/ par rapport à $V_2$/a/ est trouvée pour tous les groupes d'âge et origines (et les deux sexes). Les interactions montrent que l'effet contextuel sur F2 (comme vu pour F1) varie en fonction des groupes d'âge, avec une augmentation dont le degré diminue globalement avec l'âge (+248Hz en moyenne pour le Gr1 et +178Hz pour le Gr3), mais d'avantage dans certaines variétés régionales que d'autres. Ainsi, comme illustré sur la figure 2, la différence entre $V_2$/i/ et $V_2$/a/ tend à diminuer chez les CH et FR (pour les deux sexes) et les femmes QC entre le groupe le plus âgé comparé aux deux autres, alors que pour les BE un effet d'âge est moins clair.

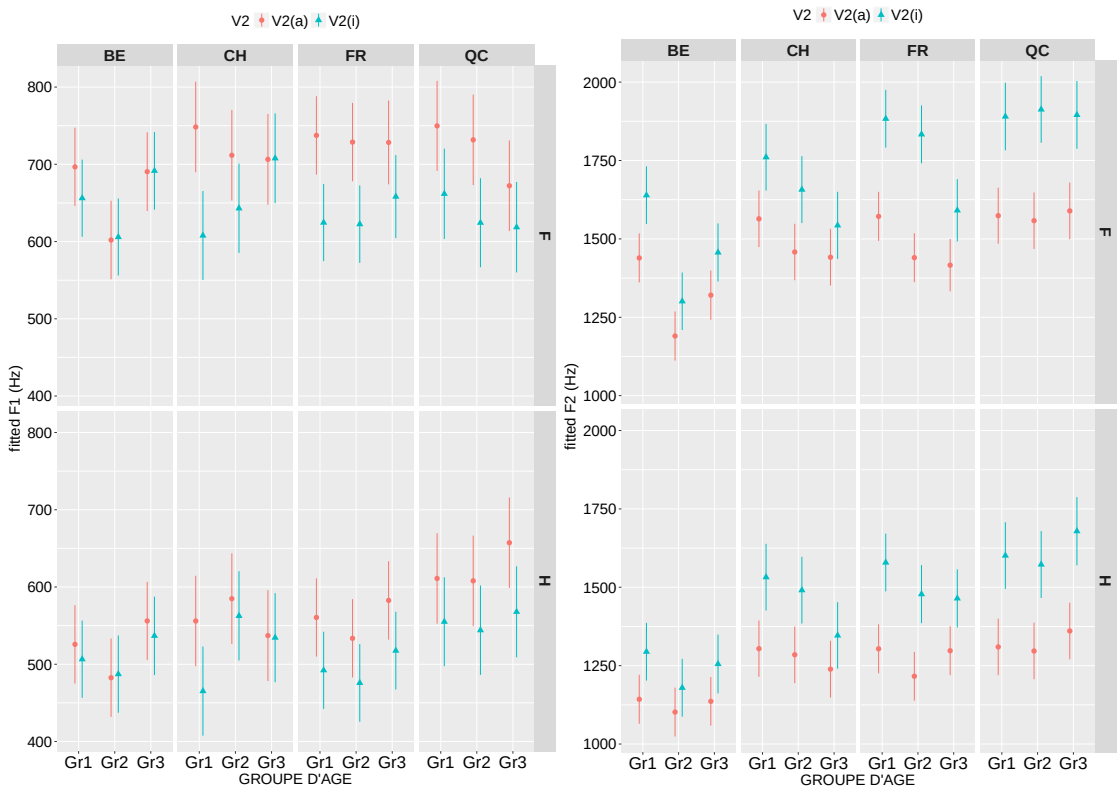


FIGURE 2 : Effets de V2 sur F1 (à gauche) et F2 (à droite) de V1 en fonction du groupe d'âge, de l'origine du locuteur et du sexe.

Enfin, concernant la durée de la voyelle $V_1$, un effet significatif est trouvé pour tous les facteurs fixes à part le SEXE, ainsi qu'une interaction $V_2$:ORIGINE. Les voyelles $V_1$ sont globalement plus longues lorsqu'elles sont suivies d'une $V_2$/i/ que d'une $V_2$/a/ ($t(167,2) = 5,3$, $p<.0001$), mais cet effet dépend de l'origine des locuteurs et ne s'avère significatif que chez les BE et les QC. D'autre part, la durée de $V_1$ est globalement plus courte chez les locuteurs du premier groupe (20-39) que chez les plus âgés : 50-69 ($t(74,8)=-7$, $p<.0001$), et 70-89 ($t(175)=-4,8$, $p<.0001$) et chez les locuteurs BE et FR, que chez les CH et QC. Afin d'observer la relation entre degré de coarticulation et durée des voyelles, une différence ($V_2$/a-i/) a été calculée par locuteur entre la moyenne des $V_1$ suivies de $V_2$/a/ et la moyenne des $V_1$ suivies de $V_2$/i/ pour F1 et pour F2. Ces différences donnent une estimation de la coarticulation par locuteur et celles-ci ont été mises en relation avec la durée moyenne des voyelles du même locuteur (toute $V_2$ confondues). La figure 3 illustre l'absence de corrélation globale entre ces deux dimensions ($r^2=0,08$ pour F2 et $r^2=-0,04$ pour F1) : les locuteurs avec des voyelles les plus longues n'ont pas moins de différences entre les deux contextes (i.e. moins d'effet contextuel), même dans le groupe des locuteurs les plus âgés (Gr3).

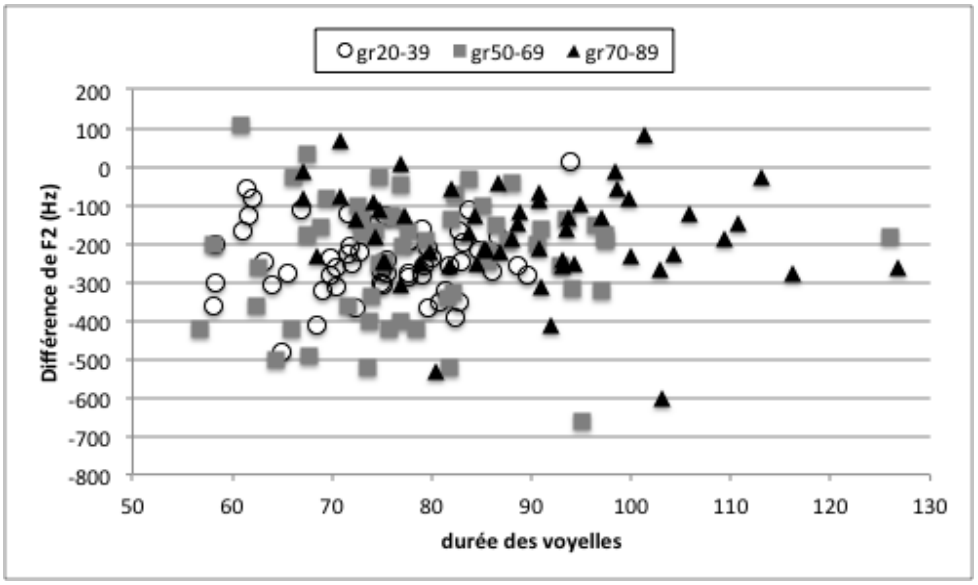



FIGURE 3 : Différence moyenne de F2 entre $V_2$/a/ et $V_2$/i/ par locuteur en fonction de la durée moyenne de ses voyelles (tous $V_2$ confondues).

# 4 Discussion

Dans cette étude, nous avons examiné les productions de locuteurs de quatre variétés régionales de français et couvrant trois classes d'âge (20-39, 50-69, 70-89), pour tester si une variation dans la coarticulation V-à-V peut être attribuée au vieillissement. Cette variation a été aussi examinée en rapport avec la durée de la voyelle et la variété régionale du locuteur.

Les résultats montrent une différence significative dans la réalisation de /a/ en fonction du contexte pour tous les groupes d'âge : l'anticipation d'un /i/ suivant abaisse le F1 et élève le F2 de /a/. Toutefois, malgré des différences régionales, on observe une diminution du degré de coarticulation avec l'âge, surtout chez les 70-89. Les locuteurs après 50 ans montrent aussi un allongement des voyelles. Les deux phénomènes de réduction de la coarticulation et d'allongement des voyelles chez

les personnes âgées peuvent-ils être expliqués par les modifications observées sur d'autres mouvements avec l'âge? Si on considère la variation dans la coarticulation et dans la durée des voyelles comme le résultat de modification au niveau du control moteur, on pourrait lier ces phénomènes à un ralentissement des mouvements et à une altération de la coordination entre mouvements qui ont été associés au vieillissement. Moins de coarticulation et un ralentissement des mouvements vocaliques pourraient être attribués à une diminution de la précision en vitesse, qui a été avancée comme une des raisons du ralentissement des mouvements chez les adultes âgés : en se focalisant sur la production correcte d'une cible acoustique, il y aurait un ralentissement dans l'articulation des segments et une diminution de la coarticulation. Une autre explication serait que la moindre coarticulation traduirait une moins bonne coordination entre mouvements et donc un encodage plus séquentiel. Une variabilité majorée dans les mouvements a été aussi associé au vieillissement, et il est également possible qu'un accroissement de la variabilité générale des cibles vocaliques (non testée ici) noie la variabilité due au contexte. Enfin il est tout à fait possible que les capacités d'anticipation des mouvements à venir ou des unités linguistiques à venir changent aussi avec l'âge : cette possibilité a été explorée dans la littérature en relation aux changements dans la coarticulation de l'enfance à l'âge adulte. Ces différentes hypothèses devront être testées à l'avenir.

La coarticulation, et donc sa réduction avec l'âge, dépendent de la variété régionale. Un effet de l'âge est clairement visible chez les locuteurs parisiens et suisses et chez les femmes québécoises. Pour les locuteurs belges par contre, il n'y a pas d'effet contextuel sur F1 et peu d'effet sur F2. Cette moindre coarticulation et l'absence de diminution avec l'âge pourraient s'expliquer par le fait que dans ces groupes de locuteurs, notamment dans les groupes plus âgés, de nombreux locuteurs utilisent une variante postérieure, [ɑ], avec un F2 bien plus bas que dans les autres régions. Une analyse plus poussée des 48 locuteurs BE est donc nécessaire. De la même façon, chez les québécois, on est en présence d'une variante beaucoup plus antérieure de type [æ] marquée par un F2 plus élevé. Outre des cibles de $V_1$ régionalement marquées, des différences dans le degré de coarticulation V-à-V en fonction des variétés régionales apparaissent. Une différence d'influence contextuelle de $V_2$ sur $V_1$ en français a déjà été montrée dans des cas d'harmonie vocalique, i.e. un cas particulier de coarticulation V-à-V qui a été phonologisée en français. Nguyen et Fagyal (2008) ont montré que le degré d'harmonie vocalique diffère entre le nord et le sud de la France : si une telle assimilation régressive est présente chez tous les locuteurs dans une certaine mesure, celle-ci est plus grande chez les locuteurs parisiens que chez les locuteurs du sud. A notre connaissance, il n'existe pas d'autres études sur les effets de la variété régionale sur la coarticulation.

Enfin, les résultats montrent que la variété régionale affecte aussi la durée des voyelles. En effet, la durée de la voyelle est plus longue chez les locuteurs suisses et québécois, par rapport aux autres groupes. En outre, chez les locuteurs belges et québécois, la voyelle /a/ est significativement plus longue lorsqu'elle est suivie par un /i/ que par un /a/. Ces variations de durée selon le contexte pourraient être mises en relation avec les variations dans la coarticulation. En fait, les deux variétés régionales qui montrent le moins d'effet coarticulatoire, les belges et les québécois (pour les hommes) sont celles qui montrent un allongement de /a/ quand elle est suivie par /i/. Cet allongement pourrait contrecarrer l'assimilation de la voyelle, en permettant à la cible /a/ d'être atteinte.

# Remerciements

Ce travail est partiellement financé par le Labex EFL (ANR-10-LABX-0083) et le projet Sinergia du FNS MoSpeeDi. Nous remercions nos collègues du projet MonPaGe pour le partage des données et F. Ivent pour ses conseils statistiques.

# Références


BARBIER, G., PERRIER, P., MÉNARD, L., PAYAN, Y., TIEDE, M., & PERKELL, J. (2017). Speech motor control in 4-year-old children versus adults: anticipation as an index of speech motor control maturity. In *7th International Conference on Speech Motor Control*.

BILODEAU-MERCURE, M., & TREMBLAY, P. (2016). Age differences in sequential speech production: articulatory and physiological factors. *Journal of the American Geriatrics Society*, 64(11).

BROWN, S. H. (1996). Control of simple arm movements in the elderly. In Advances in psychology 114, 27-52

FOUGERON C., DELVAUX V., MÉNARD L., LAGANARO M. (2018) The MonPaGe_HA Database for the Documentation of Spoken French Throughout Adulthood, *Actes de LREC 2018*.

HARRINGTON, J., PALETHORPE, S., & WATSON, C. I. (2007). Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers. In *Eighth Annual Conference of the International Speech Communication Association.*

JACEWICZ, E., FOX, R. A., WEI, L. (2010) Between-speaker and within-speaker variation in speech tempo of American English. *JASA* 128(2), 839-850.

KETCHAM, C. J., & STELMACH, G. E. (2004). Movement Control in the Older Adult. *Technology For Adaptive Aging*, 64.

MOK, P. P. (2011). Effects of vowel duration and vowel quality on vowel-to-vowel coarticulation. *Language and speech*, 54(4), 527-545.

MORRIS, R. J., & BROWN JR, W. S. (1994). Age-related differences in speech variability among women. Journal of Communication Disorders, 27(1), 49-64.

NGUYEN, N., & FAGYAL, Z. (2008). Acoustic aspects of vowel harmony in French. *Journal of Phonetics*, 36(1), 1-27.

NITTROUER, S., STUDDERT-KENNEDY, M., & NEELY, S. T. (1996). How children learn to organize their speech gestures: Further evidence from fricative-vowel syllables. *JSLHR*, 39(2), 379-389.

NOIRAY, A., MÉNARD, L., & ISKAROUS, K. (2013). The development of motor synergies in children: Ultrasound and acoustic measurements. JASA, 133(1): 444-52.

R CORE TEAM (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

RECASENS, D. (1989). Long range coarticulation effects for tongue dorsum contact in VCVCV sequences. *Speech Communication*, 8(4), 293-307.

SERENO, J. A., BAUM, S. R., MAREAN, G. C., & LIEBERMAN, P. (1987). Acoustic analyses and perceptual data on anticipatory labial coarticulation in adults and children. *JASA*, 81(2), 512-519.

SMITH, B. L., WASOWICZ, J., & PRESTON, J. (1987). Temporal characteristics of the speech of normal elderly adults. *JSLHR, 30*(4), 522-529.

TORRE III, P., & BARLOW, J. A. (2009). Age-related changes in acoustic characteristics of adult speech. *J Lang Commun Disord*, 42(5), 324-333.

VERHOEVEN, J., DE PAUW, G., KLOOTS, H. (2004) Speech rate in a pluricentric language: A comparison between Dutch in Belgium and the Netherlands. *Language & speech*, 2004, 47(3), 297-308.

CONTRERAS-VIDAL, J. L., TEULINGS, H. L., STELMACH, G. E. (1998). Elderly subjects are impaired in spatial coordination in fine motor control. *Acta psychologica,* , 100, 1-2, 25-35.

VOUSDEN, J. I., & MAYLOR, E. A. (2006). Speech errors across the lifespan. *Language and Cognitive Processes*, 21(1-3), 48-77.

WHALEN, D. H. (1990). Coarticulation is largely planned. *Journal of Phonetics*, 18, 3-35.

XUE, S. A., & HAO, G. J. (2003). Changes in the human vocal tract due to aging and the acoustic correlates of speech production: a pilot study. *JSLHR* 46(3), 689-701.

ZHARKOVA, N., HEWLETT, N., & HARDCASTLE, W. J. (2011). Coarticulation as an indicator of speech motor control development in children: An ultrasound study. *Motor Control*, 15(1), 118-140.

ZHARKOVA, N., HEWLETT, N., & HARDCASTLE, W. J. (2012). An ultrasound study of lingual coarticulation in/s V/syllables produced by adults and typically developing children. *JIPA* 42(2), 193-208.

# Voix et sélection sexuelle : une approche interdisciplinaire

Alexandre Suire[1], Michel Raymond[1] & Melissa Barkat-Defradas[1]

(1) Biologie Evolutive Humaine, Institut des Sciences de l'Evolution de Montpellier, Place Eugène Bataillon, 34095 Montpellier, France
alexandre.suire@umontpellier.fr

## RESUME

Au-delà du contenu linguistique de la parole, la voix humaine contient de nombreuses informations biopsychosociales reflétant des traits de personnalité jugées cruciaux lorsque considérés dans les contextes de séduction et de compétition humaines. Il apparait ainsi que la sélection sexuelle a contribué à favoriser la manipulation volontaire d'un certain nombre de paramètres vocaux et ce, afin de transmettre des informations non linguistiques liées à la taille, à l'attractivité et/ou à la dominance. Toutefois, la détermination des caractéristiques vocales volontairement manipulables qui confèrent un avantage évolutif, notamment en termes d'accès à des partenaires sexuels, reste une thématique peu étudiée. La présente étude vise à apporter des éléments de réponse à cette question en étudiant les comportements vocaux d'hommes et de femmes en contextes de séduction (i.e. compétition intersexuelle) et d'affrontement (i.e. compétition intrasexuelle), afin d'identifier les critères acoustiques pertinents permettant de prédire le succès d'accouplement des individus.

## ABSTRACT

Beyond the linguistic content conveyed by articulated speech, the human voice contains a lot of biological, psychological and sociological information reflecting various personality traits of the speaker that are crucial in the contexts of seduction (intersexual competition) and competition (intersexual competition). It appears that sexual selection in humans has favored the ability to manipulate volitionally a certain number of acoustic and prosodic parameters so as to convey information about the speaker's size, attractiveness and/or dominance. Nevertheless, studies investigating the evolutionary benefits (especially in terms of mating success) of vocal modulation are rather scarce. The present study aims at studying vocal behaviors in seductive vs. and competitive interactions in order to determine the vocal parameters that best predict the individuals' mating success.

MOTS-CLES: Voix; sélection sexuelle; choix de partenaire; compétition intrasexuelle; évolution humaine.

KEYWORDS: Voice; sexual selection; mate choice; intrasexual competition; human evolution.

# 1 Introduction

Au-delà du contenu linguistique qu'elle véhicule, l'une des caractéristiques fondamentale de la communication humaine est qu'elle repose essentiellement sur la voix. La qualité vocale permet en effet d'exprimer une infinité de nuances émotionnelles et affectives, allant de la douleur à la passion, de l'ivresse à la fureur, de l'exaltation à l'épouvante, etc. Hommes et femmes peuvent ainsi moduler


10.21437/JEP.2018-48

leur voix en fonction du contexte social, de leurs interlocuteurs et/ou de leurs intentions afin de transmettre toute une palette d'impressions auditives. La sélection sexuelle, mécanisme favorisant les traits biologiques conférant un avantage pour la reproduction, offre un cadre théorique pertinent pour comprendre le rôle fonctionnel de la voix humaine dans une perspective évolutive. Dans ce contexte, un grand nombre de travaux ont montré que la voix diffuse également (et au delà du message linguistique), un éventail d'information biologiques et psycho-sociales telles que, entre autres, le sexe (Puts, Gaulin et Verdolini 2006), l'âge (Linville & Fisher 1985), l'orientation sexuelle (Munson *et al.* 2006), les niveaux hormonaux (Dabbs & Mallinger 1999), la force physique (Sell *et al.* 2010) et la configuration corporelle (Hughes, Dispenza & Gallup 2004). De telles informations, transmises par des caractéristiques acoustiques et prosodiques spécifiques pouvant être manipulées, sont cruciales dès lors qu'il s'agit d'évaluer la qualité phénotypique des potentiels partenaires sexuelles et des compétiteurs.

Néanmoins, les études consacrées aux avantages directs de la modulation vocale en fonction du contexte social sont plutôt rares. Apicella, Feinberg et Marlowe (2007) ont montré que le ton de la voix des chasseurs-cueilleurs pouvait prédire de manière fiable leur succès reproducteur (i.e. un F0 bas étant corrélé à un plus grand nombre d'enfants). De même, Hodges-Simeon, Gaulin & Puts (2011) ont montré que les hommes qui s'expriment de manière plus monotone (valeurs de F0-SD moindres) rapportent plus de partenaires sexuels. Toutefois, cette étude ne s'est pas intéressée à l'étude du succès d'accouplement des femmes en relation avec leurs caractéristiques vocales. A notre connaissance, seule l'étude de (Hughes, Dispenza & Gallup 2004) a abordé la question en mesurant l'effet des caractéristiques vocales des femmes sur leur succès d'accouplement, leur nombre déclaré de rapports sexuels extra-conjugaux et l'âge de leur premier rapport sexuel. À l'exception de l'étude conduite en parole spontanée par Hodges-Simeon, Gaulin et Puts (2011), la principale limite des travaux précédemment cités est que les paramètres vocaux sont mesurés à partir d'indices vocaux statiques et peu écologiques (i.e. voyelles isolées, parole lue et non contextualisée) ce qui ne reflète évidemment pas la façon dont un individu se comporte vocalement dans le contexte d'interactions sociales réelles. En outre, nous constatons que, d'une façon générale, un intérêt tout particulier est accordé à la hauteur vocale (F0) et à ses fréquences de résonance alors même que de nombreux autres paramètres acoustiques sont perceptibles et susceptibles de varier en fonction du contexte social.

Le but de cette étude est donc de déterminer les caractéristiques acoustiques et prosodiques de la voix permettant de prédire au mieux le succès d'accouplement d'hommes et de femmes en étudiant l'évolution de la qualité vocale en parole spontanée dans deux contextes (i) un contexte de compétition (où deux individus du même sexe doivent concourir pour un(e) même partenaire), et (ii) en contexte de séduction (où un individu doit concourir pour obtenir la faveur d'un(e) partenaire de sexe opposé). L'originalité de notre travail réside dans le fait que nous avons retenu un ensemble de variables acoustiques et prosodiques peu étudié de façon conjointe dans les études précédemment citées.

# 2 Matériel et Méthode

68 femmes (âge moyen = 22,9 ans, tranche d'âge = 19 - 36 ans) et 56 hommes (âge moyen = 23 ans, tranche d'âge = 18 ans et 33 ans) ont été recrutés par le biais des réseaux sociaux et par la distribution de tracts sur le campus universitaire et autres lieux publics de Montpellier (France). Tous les participants se sont déclarés hétérosexuels et francophones et ont reçu une compensation financière pour leur participation. À leur arrivée au laboratoire, les participants ont été installés dans une chambre anéchoïque, équipée d'un microphone SennheiserTM BF 515 connecté à un ordinateur placé dans une pièce attenante. Tous les enregistrements ont été encodés à l'aide du programme

logiciel Adobe© Audition CS6 à un taux d'échantillonnage de 44 kHz - 32 bits - mono, puis sauvegardés au format .wav. Les participants ont été invités à participer à un prétendu jeu de speed-dating insipré de celui mis en œuvre par Hodges-Simeon, Gaulin & Puts (2011). La tâche consistait à gagner un rendez-vous galant avec une personne du sexe opposé tout en étant en compétition avec une autre personne du même sexe. Il a été demandé aux participants de séduire la personne du sexe opposé après avoir visionné une vidéo où un acteur et/ou une actrice professionnel.le se présentait (i.e. enregistrement du contexte de séduction). Il a ensuite été demandé au sujet d'expliquer pourquoi il.elle pensait être plus à même de gagner le rendez-vous que le.la compétiteur/compétitrice (écoute d'un enregistrement audio produit par l'acteur/actrice) après l'avoir entendu (i.e. enregistrement du contexte de compétition). A l'issue de l'expérience, les participants ont complété un questionnaire renseignant leur date et lieu de naissance, l'origine de leurs parents et grands-parents, leur statut socioéconomique (niveau d'éducation et salaire), leur statut marital (i.e. engagé ou non dans une relation) et le nombre de partenaires sexuels qu'ils ont eu au cours de la dernière année (i.e., proxy du succès d'accouplement). La variable relative au nombre de partenaires sexuels de l'année écoulée a été choisie parce qu'elle couvre une période de temps suffisamment courte pour que les souvenirs des participants à ce sujet soient exacts et parce qu'aucune modification vocale significative ne peut avoir lieu sur un intervalle temporel aussi réduit (Hodges-Siméon, Gaulin & Puts 2011). L'analyse acoustique des paramètres vocaux a été réalisée sous Praat (version 6.0.31) et l'analyse statistique sous R (version 3.4.0).

# Résultats

Afin de décorréler les variables vocales, des Analyses en Composantes Principales (ACP) ont été effectuées sur les paramètres acoustiques et prosodiques pour les deux sexes et pour chaque contexte soit sur les valeurs de F0 moyen (Hz), F0-SD (Hz, proxy de l'intonation), jitter (%, proxy de la raucité), Harmonics to Noise Ratio (HNR, proxy du souffle vocal), intensité (dB), durée (s), débit de parole (soit le temps de phonation temps de pause compris) et taux d'articulation (soit le temps de phonation hors temps de pause). Les axes retenus ont ensuite été utilisés comme variables vocales explicatives dans les analyses ultérieures. Le succès d'accouplement a été considéré comme une variable dépendante dans une régression linéaire, avec les paramètres vocaux comme variables explicatives. Comme la mesure du succès de l'accouplement consiste en un certain nombre d'événements discrets se produisant dans un intervalle de temps fixe (12 mois), une régression linéaire généralisée a été utilisée avec une structure d'erreur de quasi-Poisson compte tenu de la présence d'une légère sur-dispersion (i.e. facteur d'échelle légèrement supérieur à 1). L'âge, le statut marital et les deux cofacteurs du statut socio-économique (i.e. le revenu mensuel et le niveau d'éducation) ont été adjoints en tant que variables confondantes. Le degré de significativité de chaque terme a été évalué à partir de la comparaison du modèle excluant le terme avec le modèle incluant toutes les autres variables.

Pour les deux sexes et les deux types d'enregistrements, les trois premières dimensions ont été conservées puisqu'elles étaient suffisantes pour expliquer 60 à 70% de la variance. En règle générale, seuls les coefficients supérieurs à 0,5 et inférieurs à -0,5 ont été considérés comme déterminant principalement les axes. Pour les hommes en contexte de séduction, la première dimension est principalement déterminée par le jitter ($r$ = 074), HNR ($r$ = -0.86) et F0-SD ($r$ =0.54), la deuxième dimension par le F0 moyen ($r$ =0.84), F0-SD ($r$ = 0.57) et le taux d'articulation ($r$ = 0.54) et la troisième dimension par le débit ($r$ = 0.70) et la durée ($r$ = -0.77). En contexte de compétition, la première dimension était principalement déterminée par le jitter ($r$ = 0.63), HNR ($r$ = -0.75) et le débit ($r$ = 0.76), la deuxième dimension par la F0 moyen ($r$ = 0.74), F0-SD ($r$ = 0.77) et la durée ($r$ = 0.62), la troisième dimension uniquement par le taux d'articulation ($r$ = 0.62). Pour les femmes en contexte de séduction, la première dimension était principalement déterminée par le F0

moyen ($r$ = 0.74), le jitter ($r$ =-0.75), HNR ($r$ = 0.72) et l'intensité ($r$ = 0.57), seconde dimension par le débit ($r$ = 0.84) et le taux d'articulation ($r$ = 0.83), la troisième dimension uniquement par F0-SD ($r$ = 0.87). En contexte de compétition, la première dimension était principalement déterminée par le jitter ($r$ = -0.73), HNR ($r$ = 0.71), le F0 moyen ($r$ = 0.51) et F0-SD ($r$ = -0.53), la seconde dimension par le débit ($r$ = 0.72), le taux d'articulation ($r$ = 0.67), le jitter ($r$ = 0.54) et la durée ($r$ = -0.53), la troisième dimension par F0-SD ($r$ = 0.69) et l'intensité ($r$ = 0.60).

Quelque soit le modèle, aucune des variables contrôles n'a été significative. Dans le contexte de séduction pour les hommes, le troisième axe a un effet positif significatif sur le succès d'accouplement ($\chi2$ = 7.96, df = 1, $p < 0.05$), c'est-à-dire que les hommes qui parlent moins (temps de phonation réduit) et plus vite (débit plus rapide) ont déclaré plus de partenaires sexuels au cours de l'année écoulée. Dans le contexte de compétition, le deuxième axe a un effet négatif significatif sur le succès de l'accouplement ($\chi2$ = 4.61, df = 1, $p < 0.05$), i.e., les hommes parlant plus longtemps (temps de phonation plus important) avec une voix plus haute (F0 moyen plus élevé) et une intonation plus marqué (F0-SD plus important) ont rapporté un succès d'accouplement moindre. Dans le contexte de séduction, le premier axe a un effet négatif significatif sur le succès de l'accouplement ($\chi2$ = 4.38, df = 1, $p < 0.05$), un F0 plus haut, une qualité de voix moins rauque (valeurs de jitter augmentées et voisines de 1.97) et moins soufflée (HNR augmenté) ont rapporté moins de partenaires sexuels au cours de l'année passée. Toutefois, aucun des paramètres acoustiques n'apparaît comme significatif en contexte de compétition.

# 3 Discussion

Par cette étude, nous montrons également que l'augmentation du taux d'élocution peut être bénéfique en contexte de séduction. Ceci peut s'expliquer par le fait que plus on s'exprime rapidement plus on est jugé attrayants, convaincants et dynamiques (Street & Brady 1982). En contexte de compétition, les hommes dont les voix sont plus hautes et dont l'intonation est plus marquée (i.e. caractéristiques vocales plutôt féminines) rapportent significativement moins de partenaires sexuelles, ce qui corrobore indirectement les observations des études antérieures selon lesquelles un F0 bas et une intonation moins marquée (i.e. caractéristiques vocales plutôt masculines) est positivement corrélée au succès reproducteur (Apicella, Feinberg & Marlowe 2007 ; Hodges-Siméon, Gaulin & Puts 2011). Ces deux composantes acoustiques (i.e. voix basse et monotone) sont considérées comme des vecteurs importants de l'impression de dominance (Puts, Gaulin & Verdolini 2006) et sont également liées à la perception d'individus plus grands et plus forts (Rendall, Vokey & Nemeth 2007).

Contre toute attente, les sujets féminins attestant une voix plus haute, peu soufflée et peu rauque ont déclaré moins de partenaires sexuels au cours de la dernière année, contredisant ainsi les résultats des études précédentes conduites en anglais lesquelles ont établi de façon récurrente que des hauteurs vocales plus élevées sont jugées comme plus attractives par les membres du sexe opposé, ces dernières étant associées à des personnes plus jeunes et donc plus fertiles (Collins & Missing 2003). Toutefois, le fait que souffle vocal soit positivement liée à la jeunesse, à l'attractivité et à la féminité, probablement parce qu'il « adoucit » d'autres aspects de la parole tels que la hauteur et la résonance vocale, suggère que l'augmentation des valeurs de souffle et de raucité pourrait augmenter la probabilité du succès d'accouplement. Plusieurs études ont d'ailleurs souligné le rôle de ces deux paramètres pour l'attractivité des voix féminines (Yuasa 2010 ; Shaw & Crocker 2015).

# Références


Apicella, C. L., Feinberg, D. R., & Marlowe, F. W. (2007). Voice pitch predicts reproductive success in male hunter-gatherers. *Biology letters*, *3*(6), 682-684.

Collins, S. A., & Missing, C. (2003). Vocal and visual attractiveness are related in women. *Animal behaviour*, *65*(5), 997-1004.

Dabbs Jr, J. M., & Mallinger, A. (1999). High testosterone levels predict low voice pitchamong men. *Personality and individual differences*, *27*(4), 801-804.

Hodges-Simeon, C. R., Gaulin, S. J., & Puts, D. A. (2011). Voice correlates of mating success in men: examining "contests" versus "mate choice" modes of sexual selection. *Archives of sexual behavior*, *40*(3), 551-557.

Hughes, S. M., Dispenza, F., & Gallup, G. G. (2004). Ratings of voice attractiveness predict sexual behavior and body configuration. *Evolution and Human Behavior*, *25*(5), 295-304.

Linville, S. E., & Fisher, H. B. (1985). Acoustic characteristics of perceived versus actual vocal age in controlled phonation by adult females. *The Journal of the Acoustical Society of America*, *78*(1), 40-48.

Munson, B., McDonald, E. C., DeBoe, N. L., & White, A. R. (2006). The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech. *Journal of Phonetics*, *34*(2), 202-240.

Puts, D. A., Gaulin, S. J., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and human behavior*, *27*(4), 283-296.

Rendall, D., Vokey, J. R., & Nemeth, C. (2007). Lifting the curtain on the Wizard of Oz: biased voice-based impressions of speaker size. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(5), 1208.

Sell, A., Bryant, G. A., Cosmides, L., Tooby, J., Sznycer, D., Von Rueden, C., ... & Gurven, M. (2010). Adaptations in humans for assessing physical strength from the voice. *Proceedings of the Royal Society of London B: Biological Sciences*, rspb20100769.

Shaw, F., & Crocker, V. (2015). Creaky voice as a stylistic feature of young American female speech: an intraspeaker variation study of Scarlett Johansson. *Lifespans and Styles*, *1*(0), 21.

Street Jr, R. L., & Brady, R. M. (1982). Speech rate acceptance ranges as a function of evaluative domain, listener speech rate, and communication context. *Communications Monographs*, *49*(4), 290-308.

Yuasa, I. P. (2010). Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women?. *American Speech*, *85*(3), 315-337.

# L'abaissement de la fréquence fondamentale comme pratique de séduction

Aron Arnold
VALIBEL – Université catholique de Louvain, Place Blaise Pascal 1, 1348 Louvain-la-Neuve, Belgique
aron.arnold@uclouvain.be

## RESUME

L'objectif de cette étude est d'analyser comment la voix varie en contexte de séduction. Notre protocole de recueil de données s'inspire des scénarios de drague simulée de Puts (2005), Hodges-Simeon et al. (2010). 26 locuteurs francophones ont été enregistrés pendant qu'ils simulaient des appels téléphoniques sur des messageries vocales au cours desquels ils devaient proposer à leur correspondant une sortie au cinéma. La première tâche consistait à appeler et à inviter un ami platonique (contexte amical) et la deuxième, à inviter une personne qu'ils souhaitaient séduire (contexte de séduction). Pendant chaque tâche, trois séquences ont été produites : deux improvisations et une lecture d'un message-type adapté aux deux contextes. En comparant la parole produite en contexte amical à celle produite en contexte de séduction, nous avons constaté que les locuteurs abaissaient significativement leur fréquence fondamentale lorsqu'ils s'adressaient à une personne qu'ils souhaitaient séduire.

## ABSTRACT

**Fundamental Frequency Lowering as Seduction Practice**

The goal of this study is to analyse how voice varies in courtship context. Our data collection protocol is inspired by the mock dating scenarios designed by Puts (2005), Hodges-Simeon et al. (2010). 26 French-speakers have been recorded while simulating phone calls on voice mail systems in which they had to invite their correspondent to the cinema. During the first task they had to invite a platonic friend (friendship context), and during the second task, they had to invite a person they wanted to seduce (courtship context). During each task, three sequences where produced: two improvisations and one reading of a message adapted to the two contexts. By comparing the speech produced in friendship context to the speech produced in courtship context, we noted that the speakers lowered significantly their fundamental frequency when they spoke to someone they wanted to seduce.

MOTS-CLES : Contexte amical ; contexte de séduction ; fréquence fondamentale ; pratique sociale ; voix séduisante ; voix sexy.

KEYWORDS : Courtship context ; friendship context ; fundamental frequency ; social practice ; seductive voice ; sexy voice.


10.21437/JEP.2018-49

# 1 Introduction

Les rituels de séduction (Boetsch, Guilhem, 2005) mobilisent un ensemble de pratiques, verbales et non-verbales, à travers lesquelles les acteurs sociaux se présentent d'une certaine manière et rendent intelligibles différentes émotions, attitudes et intentions. Régulièrement, ils adoptent tout d'abord une attitude ambivalente à l'égard de leur objet de désir en alternant signaux de désir de communication et d'évitement (Givens, 1978), puis, à travers diverses marques d'attention et d'intérêt, se montrent (ou se prétendent) eux-mêmes séduits, car, comme l'écrit Baudrillard (1979, p. 112), « être séduit est bien encore la meilleure façon de séduire ». Conséquemment, la voix, qui est le vecteur par excellence pour indexer ces émotions, attitudes et intentions (Fónagy, 1983 ; Leoni, 2014), va jouer un rôle majeur dans la séduction (Ferveur, 2015).

## 1.1 Approches biologistes de l'attractivité vocale

Le rôle de la voix dans la séduction a fait l'objet de nombreuses études perceptives prenant comme cadres théoriques la psychologie évolutive et la sociobiologie (p. ex. Feinberg et al., 2005 ; Feinberg et al., 2006 ; Hughes et al., 2004 ; Pipitone, Gallup Jr, 2008 ; Puts, 2005). Les auteurs de ces études partent majoritairement du postulat que la séduction est motivée par la procréation et interprètent leurs résultats à travers ce prisme. Ils expliquent notamment régulièrement que la voix est utilisée pour évaluer la *fitness* (valeur sélective) d'un individu – elle constituerait un indice des taux d'hormones sexuelles dans son organisme et permettrait de cette manière d'estimer sa capacité à transmettre ses gènes à une descendance. Ainsi, les voix perçues comme attirantes seraient celles des « bonnes génitrices » et des « bons géniteurs » : les hommes seraient attirés par les voix de femmes aigües parce qu'elles seraient un indice de taux importants d'estrogènes et de progestérone et ainsi de fertilité ; et les femmes seraient attirées par les voix d'hommes graves parce qu'elles seraient un indice de taux élevés de testostérone, qui seraient corrélés à une grande force physique et à une plus forte compétitivité.

D'une part, il est simpliste de penser que la voix reflèterait de manière directe des taux d'hormones sexuelles et constituerait ainsi un facteur qui permettrait de sélectionner le partenaire le plus adéquat parmi un ensemble d'individus. Bien que les voix changent au cours du développement sous l'influence des changements hormonaux (Abitbol et al., 1999), d'autres facteurs, notamment génétiques (Sataloff, 1995), jouent également un rôle. Les locutrices et locuteurs héritent d'un patrimoine génétique qui influencera le développement de leur appareil phonatoire. Et cet appareil phonatoire, en fonction de sa morphologie, sera plus propice à produire des voix graves ou aigües, claires ou sombres, etc. Par ailleurs, la forme d'une voix n'est pas uniquement due à l'anatomie du locuteur, mais aussi à l'usage que celui-ci fait de cette anatomie. Un locuteur peut par exemple prendre une voix plus grave ou plus aigüe en fonction de l'identité ou de la posture qu'il souhaite indexer (Arnold, 2015 ; Fónagy, 1983), ou en fonction de la langue qu'il parle (Pépiot, Arnold 2018). Par conséquent, il est impossible d'estimer de manière fiable les taux d'hormones sexuelles de locuteurs en écoutant simplement leurs voix. D'autre part, en confondant systématiquement la séduction avec un désir de procréation, qui serait inné, propre à tout être humain, et qui déterminerait par ailleurs des préférences universelles, ces études nient la diversité des sexualités chez l'être humain, ainsi que la dimension sociale et culturellement située de la séduction (Arnold, 2016). Comme argument contre la vision biologiste portée par ces études, on peut citer le fait que des pratiques de séduction sont mobilisées dans de nombreux contextes dans lesquels la procréation n'est ni l'objectif, ni la finalité. Les scripts sexuels (Gagnon, Simon, 1973), dont la séduction fait régulièrement partie intégrante, sont extrêmement divers et le fait de prendre comme point de

référence une forme de sexualité spécifique, notamment l'hétérosexualité pénétrative et potentiellement procréative, aboutit à une marginalisation de toutes les autres formes de sexualité, comme par exemple les homosexualités, les sexualités non pénétratives, etc. (voir p. ex. Bajos, Bozon, 2016). Un autre argument contre le discours universaliste véhiculé par ces études est celui de la diversité des canons de beauté et des pratiques vocales. D'une part, les canons de beauté varient fortement d'une culture à l'autre (Reischer, Koo, 2004), et d'autre part, les pratiques vocales, comme par exemple celles relatives à la hauteur et au timbre, varient elles aussi en fonction des langues et des groupes de locuteurs (Johnson, 2006 ; Traunmüller, Eriksson, 1995). Comment alors les représentations des voix séduisantes pourraient-elles être universelles ? Des études ont montré que, bien que certaines similarités intergroupes existent, les représentations des voix séduisantes varient. Par exemple, Babel et McGuire (2013) ont trouvé des différences en étudiant les préférences de plusieurs groupes de locuteurs anglophones d'Amérique du Nord. Ils ont par exemple constaté que des locuteurs du nord de la Californie et de l'ouest du Canada trouvent plus séduisantes les voix féminines et masculines soufflées, alors que cela n'est pas le cas pour des locuteurs du Minnesota. Comme l'expliquent Boetsch et Guilhem (2005), « si les rituels de séduction sont des invariants, les formes qu'ils prennent ne le sont pas, car ils dépendent d'éléments culturellement définis qui déterminent la codification comportementale ».

## 1.2 La fréquence fondamentale en contexte de séduction

Un des paramètres acoustiques les plus analysés dans les études sur la séduction vocale est la fréquence fondamentale (F0), dont le corrélat perceptif est la hauteur (grave/aigu). Si de nombreuses études perceptives, notamment celles citées supra, concluent que les voix de femmes aigües et les voix d'hommes graves sont perçues comme étant les plus attirantes, paradoxalement, des études sur la production, comme par exemple celle menée par Tuomi et Fisher (1979), ont montré que si on demande à des locuteurs de parler avec une voix séduisante et sexy, ces derniers, quel que soit leur genre, abaissent leur F0.

Cet abaissement de F0 a donc été constaté par Tuomi et Fisher chez des locuteurs féminins et masculins. En revanche, quand Hughes, Mogilski et Harrison (2014) ont utilisé un protocole de recueil de données similaire, ils ont observé un abaissement de F0 significatif uniquement chez les sujets fémins et non pas chez les sujets masculins. Conséquemment se pose la question si l'abaissement de F0 est une pratique de séduction genrée, que l'on retrouverait plus spécifiquement chez des femmes.

Nous notons par ailleurs que les variations de la voix et de parole en contexte de séduction n'ont fait l'objet que de peu d'attention et que très peu de travaux existent sur le français. Parmi les travaux existants, nous citerons ceux sur la voix coquette de Fónagy (1983) et la voix de charme de Léon (1993). Pour combler ce manque, nous avons analysé dans la présente étude si des locutrices et des locuteurs francophones recrutés dans la région de Bruxelles et dans le Brabant wallon (Belgique) utilisent des pratiques de séduction vocales similaires à celles des locutrices et locuteurs canadiens décrits par Tuomi et Fisher (1979), et à celles des locutrices étatsuniennes décrites par Hughes, Mogilski et Harrison (2014).

# 2 Méthode

## 2.1 Participants

Les sujets qui ont participé à cette expérience – 13 femmes et 13 hommes – avaient entre 19 et 39 ans au moment de l'enregistrement du corpus. Tous étaient locuteurs de français natifs, d'origine belge ou française, et vivaient dans la région de Bruxelles ou dans le Brabant wallon (Belgique). Aucun ne présentait de trouble de la parole ou de l'audition. Les sujets ont été recrutés à travers le pool de participants de l'Université catholique de Louvain et à travers des associations étudiantes. Pour leur participation à l'étude, les sujets ont été rémunérés de 7 EUR par tranche de 30 minutes.

## 2.2 Procédure d'enregistrement et corpus

Les enregistrements se sont déroulés dans les locaux de l'Université catholique de Louvain ou au domicile des locuteurs. Les locuteurs ont été enregistrés avec un microphone serre-tête cardioïde Shure WH20XLR et un enregistreur numérique Edirol/Roland R09-HR. Les sessions d'enregistrement ont duré en moyenne une demi-heure.

Notre protocole de recueil de données s'inspire des scénarios de drague simulée que Puts (2005) et Hodges-Simeon, Gaulin et Puts (2010) avaient élaborés d'après Simpson et al. (1999). Nous avons demandé aux sujets de simuler des appels téléphoniques et de laisser des messages sur des messageries vocales de correspondants fictifs. Dans ces messages, ils devaient proposer à leurs correspondants une sortie au cinéma. Nous avons demandé aux sujets d'appeler dans un premier temps un très bon ami, et dans un deuxième temps, une personne qu'ils souhaitaient séduire, tout en utilisant une voix qu'ils qualifieraient de « séduisante » et « sexy ». Nous n'avons donné aucune autre instruction sur la manière dont ils devaient parler. Nous avons choisi le thème de la sortie au cinéma parce qu'il s'agit d'une activité qui se fait régulièrement entre amis platoniques, mais qui est aussi culturellement associée aux rendez-vous amoureux (Bogle, 2008). Les deux tâches ont toujours été réalisées dans le même ordre, dans les mêmes conditions, à environ deux minutes d'intervalle l'une de l'autre.

L'objectif de ce protocole était de collecter de la parole correspondant à deux contextes comparables mais distincts – contexte amical et contexte de séduction – afin de pouvoir isoler les caractéristiques phonétiques de la parole de séduction. La parole des sujets a été enregistrée pendant l'accomplissement de deux tâches, correspondant aux deux contextes. Chaque tâche comportait trois séquences : deux séquences d'improvisation pendant lesquelles les sujets étaient libres de formuler leurs messages comme ils le souhaitaient, puis une séquence pendant laquelle les sujets devaient reproduire le message suivant, tout en gardant les intonations utilisées pendant les improvisations :

> « Salut (prénom du correspondant). C'est (prénom du sujet). J'espère que tu vas bien ! Est-ce que ça te dirait d'aller au cinéma avec moi ce soir ? Rappelle-moi. Salut. »

Nous avons choisi ce message-type parce qu'il comporte un contenu lexical et sémantique adapté aux deux contextes étudiés, et parce qu'il oblige les sujets à indexer uniquement à travers leur voix la nature du rendez-vous – si la sortie au cinéma se fera dans le cadre d'un rendez-vous amical ou dans le cadre d'un rendez-vous amoureux.

Afin de mettre les sujets « en situation », chaque séquence a été amorcée par le message audio préenregistré suivant :

> « Messagerie Orange, bonjour ! La personne que vous essayez de joindre n'est pas disponible. Veuillez laisser votre message après le bip ».

### 2.3 Analyse des données

Notre corpus a été constitué dans le cadre d'un projet de recherche plus large sur l'érotisme vocal au cours duquel un ensemble de paramètres acoustiques et prosodiques seront étudiés. Dans le présent article nous nous limiterons à présenter des résultats intermédiaires relatifs à la F0 moyenne des 26 locuteurs enregistrés à ce jour.

Les relevés de la F0 moyenne ont été effectués manuellement à l'aide du logiciel Praat (Boersma, 2017). Pour chaque locuteur, six relevés ont été réalisés : trois correspondant au contexte amical et trois correspondant au contexte de séduction. Un ensemble de 156 relevés de F0 – 78 en contexte amical et 78 en contexte de séduction – a ensuite fait l'objet d'un test des rangs signés de Wilcoxon, afin de vérifier s'il existe des différences significatives entre les F0 de parole produite en contexte de séduction et de parole produite en contexte amical. Le choix d'un test non-paramétrique a été motivé par la distribution non normale de notre échantillon. Ce test a été réalisé sur l'ensemble des relevés, puis séparément sur les relevés correspondant aux improvisations et sur les relevés des messages-types.

## 3 Résultats

Les F0 des sujets féminins sont présentées dans le tableau n° 1 et celles de sujets masculins dans le tableau n° 2. Ces F0 sont des moyennes des trois séquences enregistrées (deux improvisations et un message-type) par contexte (contexte amical et contexte de séduction).

| Locuteur | F0 ami. | F0 séd. | Abaiss. F0 | Abaiss. DT |
|---|---|---|---|---|
| Loc F 1 | 263 | 223 | -15% | -2,9 |
| Loc F 2 | 224 | 214 | -4% | -0,8 |
| Loc F 3 | 189 | 179 | -5% | -0,9 |
| Loc F 4 | 242 | 228 | -6% | -1 |
| Loc F 5 | 231 | 174 | -25% | -4,9 |
| Loc F 6 | 209 | 204 | -3% | -0,4 |
| Loc F 7 | 298 | 231 | -23% | -4,4 |
| Loc F 8 | 243 | 220 | -9% | -1,7 |
| Loc F 9 | 245 | 212 | -13% | -2,5 |
| Loc F 10 | 261 | 242 | -8% | -1,3 |
| Loc F 11 | 230 | 233 | 1% | 0,2 |
| Loc F 12 | 192 | 174 | -10% | -1,7 |
| Loc F 13 | 200 | 194 | -3% | -0,5 |
| **Moyenne** | **233** | **210** | **-10%** | **-1,8** |

*Tableau 1 – Sujets féminins : F0 moyenne en Hz en contexte amical et contexte de séduction ; abaissement de F0 en contexte de séduction ; abaissement en demi-tons.*

| Locuteur | F0 ami. | F0 séd. | Abaiss. F0 | Abaiss. DT |
|---|---|---|---|---|
| Loc H 1 | 104 | 68 | -34% | -7,4 |
| Loc H 2 | 92 | 74 | -19% | -3,8 |
| Loc H 3 | 127 | 120 | -5% | -1 |
| Loc H 4 | 121 | 118 | -2% | -0,4 |
| Loc H 5 | 92 | 85 | -7% | -1,4 |
| Loc H 6 | 91 | 87 | -4% | -0,8 |
| Loc H 7 | 109 | 98 | -10% | -1,8 |
| Loc H 8 | 123 | 117 | -5% | -0,9 |
| Loc H 9 | 105 | 93 | -11% | -2,18 |
| Loc H 10 | 119 | 128 | 8% | 1,3 |
| Loc H 11 | 116 | 103 | -11% | -2,1 |
| Loc H 12 | 120 | 112 | -6% | -1,2 |
| Loc H 13 | 155 | 116 | -25% | -5 |
| **Moyenne** | **113** | **101** | **-11%** | **-1,9** |

*Tableau 2 – Sujets masculins : F0 moyenne en Hz en contexte amical et contexte de séduction ; abaissement de F0 en contexte de séduction, abaissement en demi-tons.*

Lorsqu'on compare les deux contextes, on peut constater que 24 des 26 sujets ont abaissé leur F0 quand ils ont produit de la parole adressée à une personne qu'ils souhaitaient séduire. L'abaissement moyen des sujets féminins est de -10%, correspondant à -1,8 demi-tons, et celui des sujets masculins de -11%, correspondant à -1,9 demi-tons. Un test des rangs signés de Wilcoxon sur les 156 relevés de F0 – 78 en contexte amical et 78 en contexte de séduction – a confirmé qu'il existe une différence significative de F0 moyenne entre les deux contextes étudiés ($z=6,7$ ; $p<0,001$). Comme on peut le voir sur la figure n° 1, les locuteurs utilisent une plage de variation dans des fréquences plus basses en contexte de séduction qu'en contexte amical.

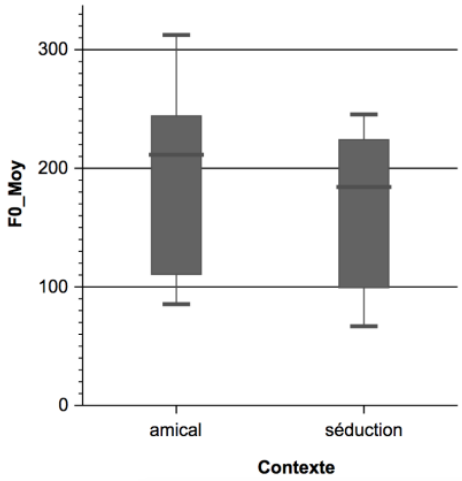



*Figure 1 : Plages de variation moyennées en Hz en contexte amical et en contexte de séduction*

Nous avons ensuite analysé les séquences improvisées et les messages-types séparément. Un test des rangs signés de Wilcoxon a montré une différence significative ($z=5,19$ ; $p<0,001$) entre séquences improvisées en contexte amical (n=52) et séquences improvisées en contexte de séduction (n=52).

De même, une différence significative (z=4,26 ; p<0,001) a été constatée entre messages-types en contexte amical (n=26) et messages-types en contexte de séduction (n=26).

# 4 Discussion

L'analyse de la F0 moyenne a montré que celle-ci est régulièrement plus basse en contexte de séduction qu'en contexte amical. C'est ce qu'illustre par exemple la figure n° 2 qui représente les courbes mélodiques d'un même locuteur masculin dans les deux contextes étudiés. La courbe rouge, qui correspond à la parole produite en contexte de séduction, se trouve dans des fréquences plus basses que la courbe bleue, qui elle correspond à la parole produite en contexte amical.

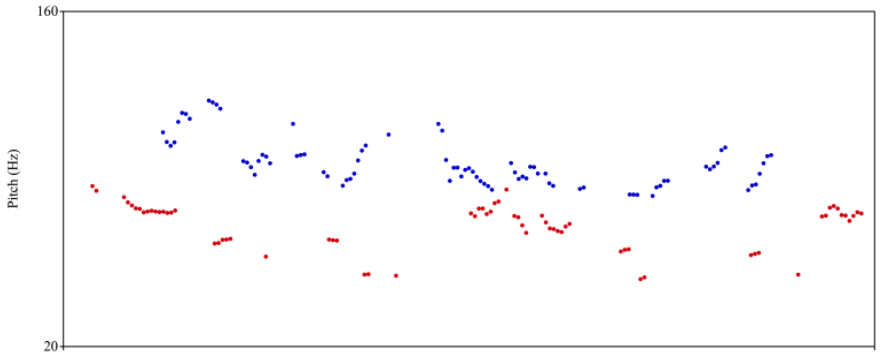



*Figure 2 – Messages-type du locuteur H 2 : F0 en contexte amical en bleu, F0 en contexte de séduction en rouge*

Nous notons que 24 des 26 sujets qui ont participé à notre étude ont abaissé leur F0 lorsqu'ils avaient comme consigne de produire une voix « séduisante » et « sexy ». Étant donné que l'action d'abaisser sa F0 en contexte de séduction a pu être observée chez la quasi-totalité des sujets (92,31%), nous interprétons cette régularité dans l'abaissement comme l'indice d'une pratique sociale – un comportement routinier, régulier, identifiable et reconnaissable (Reckwitz, 2002). Un argument qui va dans ce sens est que l'on retrouve régulièrement dans le cinéma ou dans la littérature occidentale des figures de séductrices et de séducteurs à voix graves : la voix de la femme fatale du film noir, grave et soufflée ; la voix du héros de roman d'amour, décrite comme grave et résonnante, etc. Les médias participent ainsi à véhiculer cette représentation de la voix séductrice. Il est par ailleurs intéressant de noter que le fait d'associer la voix grave à la séduction et à la sexualité n'est pas un phénomène récent. Déjà au 19ème siècle, la voix grave a souvent été décrite comme étant une des caractéristiques des prostituées, notamment par Ellis (1896) au Royaume-Uni et par Parent-Duchâtelet (1837) en France. Parent-Duchâtelet explique que pour certains physiologistes de son époque, le « caractère viril » des voix des prostitués serait dû à leur lascivité et à certaines pratiques sexuelles « que réprouve la nature » (1837, p. 197). Cette croyance en un lien entre activité sexuelle et hauteur de voix peut selon Graddol et Swann (1989, p. 17) encore être retrouvée de nos jours parmi les chanteurs d'opéra. Ils expliquent qu'au Royal Opera House de Londres, on conseille aux ténors et sopranos d'éviter les rapports sexuels avant les représentations, aux barytons de les limiter à une à deux fois par semaine, et aux basses d'en avoir tous les soirs.

Si l'on s'intéresse maintenant aux raisons de l'abaissement de F0 en contexte de séduction, on peut émettre l'hypothèse que cet abaissement est motivé par certaines attitudes ou qualités qui sont cognitivement associées aux voix graves, comme par exemple la confiance en soi ou l'assertivité (Ohala, 1994). Ces deux traits sont souvent vus comme caractérisant les figures des séductrices et des séducteurs. Des expériences perceptives que nous conduirons prochainement dans le cadre de ce projet de recherche permettront d'étudier si les voix produites en contexte de séduction sont perçues comme indexant un degré supérieur de confiance en soi et d'assertivité à celles produites en contexte amical.

Nos données n'ont pas montré de différence majeure entre sujets féminins et masculins : les deux groupes ont abaissé leur F0 en contexte de séduction. Chez les sujets féminins, nous avons mesuré un abaissement de -10% correspondant à -1,8 demi-tons, et chez les sujets masculins un abaissement de -11%, correspondant à -1,9 demi-tons. L'abaissement de F0 en contexte de séduction ne semble donc pas être une pratique liée au genre. Il est cependant possible que des différences prosodiques existent entre femmes et hommes, mais que celles-ci aient été invisibilisées par la quantification de F0 moyennes. En écoutant les enregistrements de nos locutrices et locuteurs, nous avons fait des constats similaires à ceux de Fónagy (1983) : nous avons par exemple remarqué que la parole des locutrices en contexte de séduction était fréquemment accompagnée de « glissements ultra-rapides vers le haut » (Fónagy, 1983, p. 131). Une analyse fine des contours intonatifs permettra de mieux étudier s'il existe des différences de genre dans les pratiques vocales de séduction.

# Remerciements

Nous tenons à remercier chaleureusement toutes les personnes qui ont accepté de participer à ce projet de recherche en tant que sujets, ainsi que Charlotte Kouklia (Laboratoire de phonétique et phonologie) pour son aide lors du design du protocole de recueil de données. Nous remercions également le programme MOVE-IN Louvain pour le financement de ce projet.

# Références

ABITBOL J., ABITBOL P., ABITBOL B. (1999). Sex hormones and the female voice. *Journal of Voice: Official Journal of the Voice Foundation* 13(3), 424- 446.

ARNOLD A. (2016). Idéologies de genre et construction des savoirs en sciences phonétiques. *GLAD! Revue sur le langage, le genre, les sexualités* 1.

ARNOLD A. (2015). Voix et transidentité : changer de voix pour changer de genre ? *Langage et société* 151(1), 87- 105.

BABEL M., MCGUIRE G. (2013). Perceived vocal attractiveness across dialects is similar but not uniform. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 426- 430.

BAJOS N., BOZON M. (2016). *Enquête sur la sexualité en France: Pratiques, genre et santé*. Paris: La Découverte.

BAUDRILLARD J. (1979). *De la séduction*. Paris: Galilee.

BOERSMA P., WEENINK D. (2017). Praat: doing phonetics by computer [Logiciel]. Version 6.0.36, publiée le 11 Novembre 2017 sur le site www.praat.org.

BOETSCH G., GUILHEM D. (2005). Rituels de séduction, Rituals of Seduction. *Hermès, La Revue* (43), 179- 188.

BOGLE K. A. (2008). *Hooking Up: Sex, Dating, and Relationships on Campus*. New York: NYU Press.

ELLIS H. (1896). *Man and Woman: A Study of Secondary and Tertiary Sexual Characters*. London: Walter Scott.

FEINBERG D. R., JONES B. C., DEBRUINE L. M., MOORE F. R., LAW SMITH M. J., CORNWELL R. E., … PERRETT D. I. (2005). The voice and face of woman: One ornament that signals quality? *Evolution and Human Behavior* 26(5), 398- 408.

FEINBERG D. R., JONES B. C., LAW SMITH M. J., MOORE F. R., DEBRUINE L. M., CORNWELL R. E., … PERRETT D. I. (2006). Menstrual cycle, trait estrogen level, and masculinity preferences in the human voice. *Hormones and Behavior* 49(2), 215- 222.

FERVEUR C. (2015). Les voi(x)es de la séduction, The voices of seduction. *Enfances & Psy* (68), 103- 116.

FONAGY I. (1983). *La vive voix: essais de psycho-phonétique*. Paris: Payot.

GAGNON J. H., SIMON W. (1973). *Sexual conduct: the social sources of human sexuality*. London: Aldine Pub. Co.

GIVENS D. B. (1978). The Nonverbal Basis of Attraction: Flirtation, Courtship, and Seduction. *Psychiatry* 41(4), 346- 359.

GRADDOL D., SWANN J. (1989). *Gender voices*. Hoboken: Wiley-Blackwell.

HODGES-SIMEON C. R., GAULIN S. J. C., PUTS D. A. (2010). Voice Correlates of Mating Success in Men: Examining « Contests » Versus « Mate Choice » Modes of Sexual Selection. *Archives of Sexual Behavior* 40(3), 551- 557.

HUGHES S. M., DISPENZA F., GALLUP JR G. G. (2004). Ratings of voice attractiveness predict sexual behavior and body configuration. *Evolution and Human Behavior* 25(5), 295- 304.

HUGHES S. M., MOGILSKI J. K., HARRISON M. A. (2014). The Perception and Parameters of Intentional Voice Manipulation. *Journal of Nonverbal Behavior* 38(1), 107- 127.

JOHNSON K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics* 34(4), 485- 499.

LEON P. R. (1993). *Précis de phonostylistique: parole et expressivité*. Paris: Nathan.

LEONI F. (2014). *Des Sons et des Sens. la Physionomie Acoustique des Mots*. Lyon: Ecole Normale Supérieure.

OHALA J. (1994). The frequency codes underlies the sound symbolic use of voice pitch. In L. Hinton, J. Nichols, & J. Ohala (Éd.), *Sound Symbolism* (p. 325- 347). Cambridge: Cambridge University Press.

PARENT-DUCHATELET A.-J.-B. (1837). *De la prostitution dans la ville de Paris, considérée sous le rapport de l'hygiène publique, de la morale et de l'administration*. Paris: J.-B. Baillière.

PEPIOT E., ARNOLD A. (2018). Étude des variations de fréquence fondamentale relatives au genre chez des bilingues Anglais/Français. *XXXIIe Journées d'Études sur la Parole.*

PIPITONE R. N., GALLUP JR G. G. (2008). Women's voice attractiveness varies across the menstrual cycle. *Evolution and Human Behavior* 29(4), 268- 274.

PUTS D. A. (2005). Mating context and menstrual phase affect women's preferences for male voice pitch. *Evolution and Human Behavior* 26(5), 388- 397.

RECKWITZ A. (2002). Toward a Theory of Social Practices: A Development in Culturalist Theorizing. *European Journal of Social Theory* 5(2), 243- 263.

REISCHER E., KOO K. S. (2004). The Body Beautiful: Symbolism and Agency in the Social World. *Annual Review of Anthropology* 33, 297- 317.

SATALOFF R. T. (1995). Genetics of the voice. *Journal of Voice* 9(1), 16- 19.

SIMPSON J. A., GANGESTAD S. W., CHRISTENSEN P. N., LECK K. (1999). Fluctuating asymmetry, sociosexuality, and intrasexual competitive tactics. *Journal of Personality and Social Psychology* 76(1), 159- 172.

TRAUNMÜLLER H., ERIKSSON A. (1995). *The frequency range of the voice fundamental in the speech of male and female adults*. Manuscrit: http://www2.ling.su.se/staff/hartmut/f0_m&f.pdf

TUOMI S. K., FISHER J. E. (1979). Characteristics of Simulated Sexy Voice. *Folia Phoniatrica et Logopaedica* 31(4), 242- 249.

# Évolution des habiletés articulatoires au stade du babillage : le timing des syllabes CV

Mélanie Canault[1, 2], Naomi Yamaguchi[3], Nikola Paillereau[3], Johanna-Pascale Roy[4], Christophe dos Santos[5] et Sophie Kern[1]

(1) Laboratoire Dynamique du Langage – UMR 5596 CNRS, Lyon, France
(2) Institut des Sciences et Techniques de la Réadaptation, Lyon, France
(3) Laboratoire de Phonétique et Phonologie – UMR 7018 CNRS, Paris, France
(4) Département de Langues, linguistique et traduction, Univ. Laval, Qc, Canada
(5) IBrain – UMR 1253, Université de Tours, Inserm, Tours, France
melanie.canault@univ-lyon1.fr, naomi.yamaguchi@univ-paris3.fr, nikola.paillereau@mac.com, Johanna-Pascale.Roy@lli.ulaval.ca, christophe.dossantos@univ-tours.fr, Sophie.Kern@cnrs.fr

## RÉSUMÉ

Au cours du processus d'acquisition du langage, le babillage correspond au stade de l'émergence des syllabes. C'est un stade au cours duquel le potentiel articulatoire commence à se construire. Le timing de l'exécution des gestes articulatoires pourrait témoigner du développement des habiletés motrices. Ce phénomène peut être inféré à partir de l'observation de la durée syllabique et de sa variabilité. Les syllabes de type CV de 22 enfants nés à terme, enregistrés mensuellement, entre 8 mois et 14 mois, ont été analysées. Une diminution et une stabilisation de la durée de la syllabe ressort des résultats. Des changements majeurs dans l'organisation temporelle syllabique apparaissent autour de 10-11 mois. Les syllabes sans changement de position linguale entre la consonne et la voyelle, pourraient présenter des marques de contrôle plus précoces que les autres.

## ABSTRACT

**Articulatory abilities evolution at babbling stage: the timing of CV syllables**

During language acquisition process, babbling is associated with the emergence of syllables. This stage is characteristic of articulatory abilities development, which can be described in terms of the timing of articulatory gesture production. Such timing can be inferred from the observation of the syllabic duration and its variability. The CV syllables of 22 toddlers, monthly recorded between 8 and 14 months of age, were analyzed. A decrease and a stabilization of the syllabic duration have been observed. Major changes in syllabic timing appear around 10-11 months. Syllables with no change in lingual position between the consonant and the vowel, seem to be the first to gain control.

MOTS-CLÉS : babillage, habiletés oro-motrices, syllabe, patrons de cooccurrence, durée, variabilité temporelle

KEYWORDS: babbling, oromotor skills, syllable, co-occurrence patterns, duration, temporal variability


10.21437/JEP.2018-50

# 1 Introduction

Le babillage est une étape du processus d'acquisition du langage qui se caractérise par l'émergence des syllabes. Chez la majorité des enfants, il apparait brusquement vers l'âge de 5-6 mois. Il est alors considéré comme rudimentaire, car les syllabes relèvent d'une articulation lâche et de transitions lentes (Oller, 1980). Quand les articulations se raffermissent, le bébé entre dans le babillage canonique. Celui-ci s'établit entre l'âge de 6-8 mois et celui de 12 mois. Pour certains, cette période se subdivise en deux stades successifs : celui du babillage dupliqué qui se définit par la répétition de la même syllabe au cours d'un énoncé, puis celui du babillage varié qui correspond à l'enchainement de syllabes différentes au sein d'un énoncé (Oller, 1980). D'autres, en revanche, défendent la simultanéité des séquences dupliquées et variées, mais en des proportions différentes au cours du temps (Davis, MacNeilage, 2000 ; Lipkind et al., 2013). Le babillage perdure jusqu'à l'émergence des premiers mots avec lesquels il coexiste à 12 mois. C'est la période du babillage mixte. Puis, il diminue progressivement pour disparaitre aux environs de 18 mois.

Même si les manifestations du babillage sont concernées par une importante variabilité interindividuelle, certaines de ses caractéristiques, telles que l'âge d'apparition, la quantité et la complexité phonétique des productions, ou encore l'organisation temporelle, apparaissent comme des indicateurs précoces d'un développement atypique, voire comme des prédicteurs des développements langagiers ultérieurs. Le babillage est donc déterminant pour le développement du langage, d'autant plus que le potentiel articulatoire va considérablement se développer au cours de cette période.

## 1.1 Immaturité motrice et dominance mandibulaire précoces

Les syllabes précoces du babillage sont la conséquence de la superposition de la phonation au mouvement rythmique d'élévation et d'abaissement de la mandibule. Ce mouvement mandibulaire constitue le cadre de la parole sur lequel viendra se superposer le contenu, c'est-à-dire le déplacement des autres articulateurs (MacNeilage, 1998). La mandibule est donc fortement impliquée dans les productions orales du babillage précoce et serait même le seul articulateur engagé (Munhall, Jones, 1998 ; Green et al., 2002). Toutefois, même si le mouvement mandibulaire constitue le geste de base permettant l'émergence de la syllabe, celui-ci n'est pas encore contrôlé. Ce manque d'efficience est notamment observable à travers les déplacements lents (Nip et al., 2009, 2011) et variables (Steeve et al., 2008) des articulateurs. Le rythme de production de la parole adulte avoisine 5-6Hz (Pellegrino et al., 2011) alors que celui du jeune enfant (8-16 mois) se situe autour de 3Hz (Dolata, 2008 ; Canault et al., 2011). Les faibles habiletés articulatoires constatées à ce stade transparaitraient également dans les patrons d'association Consonne-Voyelle. Les mouvements linguaux seraient, en effet, fortement dépendants de ceux de la mandibule. Cela signifierait que lors de la réalisation d'une syllabe, les déplacements de la langue interviendraient essentiellement sur le plan vertical sous l'impulsion de l'oscillation mandibulaire. Autrement dit, il existerait une certaine inertie de la langue dans la dimension horizontale au cours de la réalisation d'une syllabe et la position linguale initiée pour la consonne serait maintenue pour la production de la voyelle. Si la langue est en position de repos, une voyelle centrale serait associée à une consonne bilabiale (ex : [ba]). En revanche, quand la langue réalise un léger mouvement antérieur, l'association d'une consonne coronale à une voyelle d'avant serait produite (ex : [dɛ]) ; et lorsque la langue recule dans la cavité buccale, une consonne vélaire serait combinée à une voyelle postérieure (ex : [gu]) (Davis,

MacNeilage, 1995, 1998, 2000). Ces patrons associatifs sont dits préférentiels et sont observés dans les productions enfantines de nombreuses langues du monde (Vihman, 1992).

### 1.2 Émergence du contrôle articulatoire

Le potentiel articulatoire va néanmoins progresser au cours de la période du babillage. Les gestes de la langue, articulateur porté et contraint par la mandibule, commenceraient à se dissocier de ceux imposés par le cadre vertical de l'oscillation mandibulaire. De ce fait, entre l'âge de 8 mois et celui de 12 mois, une augmentation des mouvements horizontaux de la langue associés à des déplacements verticaux de la mandibule émergerait (Canault et al., 2008). De plus, à l'âge d'1 an, les patrons de déplacement de la mandibule seraient les premiers à se rapprocher de ceux de l'adulte pour l'activité de parole tant sur le plan de leur trajectoire que de leur stabilité (Green et al., 2002). Une accélération (Green, Wilson, 2006 ; Nip et al., 2009) et une stabilisation (Green et al., 2002) des mouvements mandibulaires sembleraient notamment s'opérer au cours de cette période développementale. Cependant, cette évolution ne serait pas linéaire (Studdert-Kennedy et al., 1991 ; Smith, Thelen, 2003 ; Green et al., 2010). Une période critique serait observée autour de 10-11 mois au cours de laquelle l'exécution de la syllabe s'accélèrerait tout en présentant une forte variabilité temporelle (Canault et al., 2011).

La syllabe, conséquence de l'oscillation mandibulaire et champ d'application de la coarticulation, est donc une unité intéressante pour l'observation du développement des habiletés oro-motrices chez le jeune enfant. L'objectif de ce travail est ainsi de rendre compte de l'émergence du contrôle articulatoire au stade du babillage grâce, à un important échantillon de données recueilli pour le français. L'accélération et la stabilisation de l'exécution des gestes articulatoires pouvant témoigner du développement des habiletés motrices, il s'agira d'observer la durée syllabique et sa variabilité entre l'âge de 8 mois et celui de 14 mois. Nous comparerons également le comportement temporel des syllabes impliquant des associations préférentielles aux non préférentielles, afin de tester l'existence des profils d'évolution temporelle différents pour ces deux types d'associations. Les syllabes intégrant des patrons de cooccurrence préférentiels, c'est-à-dire sans changement de position linguale entre la consonne et la voyelle, pourraient présenter des marques de contrôle plus précoces.

## 2 Méthodologie

### 2.1 Population

22 enfants nés à terme, d'environnement monolingue francophone sans pathologie et/ou handicap connu(s) ont ainsi été recrutés (11 garçons et 11 filles). Les jumeaux ainsi que les enfants présentant une prématurité, une pathologie ORL, un handicap neurologique, un syndrome de dysoralité sensorielle et un trouble mental connu n'ont pas été inclus à l'expérimentation.

### 2.2 Recueil de données

Tous les représentants légaux des participants ont signé un formulaire de consentement avant le début des expérimentations. Chaque enfant a été enregistré mensuellement (respectant un délai

maximum de 15 jours après la date « anniversaire ») de l'âge de 8 mois à l'âge de 14 mois, soit 7 sessions par enfant. Au cours de chaque séance, trois types de données ont été recueillies :

1. Un questionnaire sur l'oralité alimentaire de l'enfant pour écarter l'absence de dysoralité sensorielle ou son apparition.
2. Un questionnaire informatisé sur le développement communicatif de l'enfant (IFDC, Kern, Gayraud, 2010) afin de vérifier la compétence de l'enfant cible par rapport à la norme.
3. Un enregistrement des productions orales de l'enfant. Les enregistrements avaient lieu soit sur le lieu de garde de l'enfant soit au domicile des parents, dans une pièce suffisamment calme ou isolée pour éviter la superposition de bruits parasites. La durée moyenne des enregistrements est d'une heure. Pour réaliser l'enregistrement audio, un dispositif mobile (Zoom® Handy Recorder H1) était utilisé. Un minimum de 40 syllabes par session d'enregistrement et par enfant était attendu. Plusieurs sessions ont parfois été nécessaires pour atteindre cet objectif.

### 2.3 Segmentation et annotation

Les productions orales des enfants cibles ont été extraites (élimination des chevauchements, cris, vocalisations et trilles) et segmentées en syllabes, puis transcrites à l'aide du logiciel Praat®. La segmentation et la transcription ont été effectuées par des experts et/ou des personnes formées spécifiquement en suivant le même protocole. Un contrôle de la fiabilité inter juge est en cours. Des concertations systématiques entre plusieurs experts ont été mises en place dans le cas des séquences de production jugées ambiguës. À partir de ces informations, la durée de chaque syllabe a pu être obtenue grâce à l'adaptation d'un Script Praat (# Copyright 12.3.2002 Mietta Lennes) et les caractéristiques articulatoires des constituants de la syllabe ont pu être extraites automatiquement grâce à un algorithme adapté de Yamaguchi et al. (2015).

### 2.4 Analyses

16 268 syllabes ont ainsi été annotées. Les syllabes inférieures à 100 ms et celles supérieures à 1000 ms ont été exclues, soit 4,9% des syllabes de l'échantillon initial. Sur les 15 472 syllabes restantes, nous avons retenu les syllabes de type CV pour cette étude, c'est-à-dire 11 261 syllabes soit 72,7% de l'échantillon analysable. Les différentes analyses statistiques ont été réalisées avec le logiciel R. Des tests de $\chi^2$ d'indépendance ont permis de vérifier l'influence de l'âge sur l'évolution des proportions d'associations produites (préférentielles ou non préférentielles). Des ANOVAs à mesures répétées ont été réalisées afin de rendre compte de l'évolution du timing de la syllabe (durée syllabique moyenne et écart type moyen) en fonction de l'âge et du type d'associations produites.

## 3 Résultats

### 3.1 Le corpus

Si l'on additionne l'ensemble des occurrences correspondant à des associations non préférentielles, ce type d'associations semble prédominer à tous les stades (Fig. 1). Le test d'indépendance du $\chi^2$ montre que les variables de l'âge et de la proportion des associations préférentielles et non préférentielles ne sont pas indépendantes ($\chi^2(6) = 60{,}267$, $p = 3.972e\text{-}11$). Autrement dit, l'âge influence la proportion du type d'associations et cet effet est perceptible à 10 mois. En effet, le

pourcentage des associations préférentielles et non préférentielles est relativement stable sur l'ensemble de la période d'observation, excepté à cet âge où la dominance des associations non préférentielles s'intensifie. 36,1% des syllabes CV relèvent de patrons de cooccurrence préférentiels et 63,8% de patrons de cooccurrence non préférentiels à cet âge.

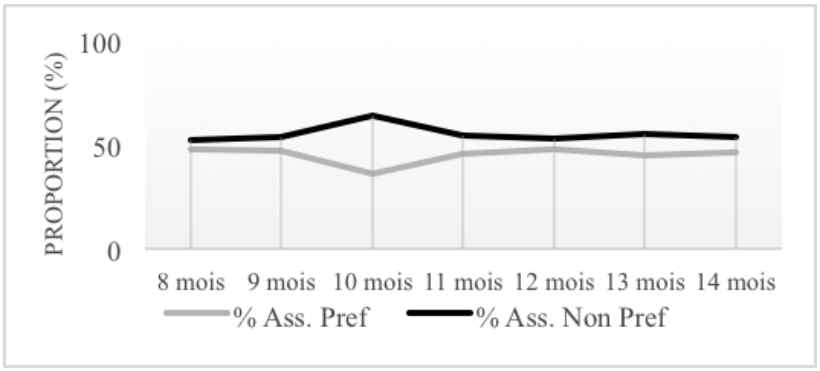



FIGURE 1 : Proportion des associations préférentielles *vs* non préférentielles entre l'âge de 8 mois et celui de 14 mois

## 3.2 Évolution du timing (durée et variabilité)

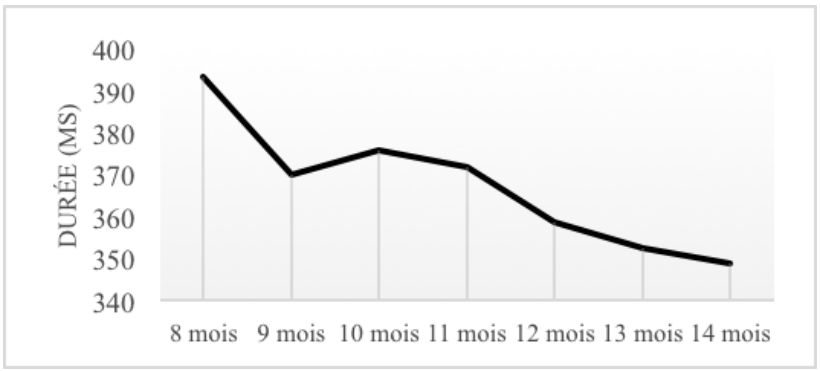



FIGURE 2 : Durée syllabique (ms) entre l'âge de 8 mois et celui de 14 mois

La durée des syllabes CV diminue significativement avec l'âge ($p = 0,031$). Cette diminution n'est pas linéaire avant l'âge de 10 mois (Fig. 2). La variabilité temporelle, mesurée par l'écart-type moyen des durées, présente également une tendance à la diminution entre 8 mois et 14 mois ($p = 0,057$). L'écart-type moyen tend à augmenter entre 8 mois et 11 mois, puis diminue entre 11 mois et 14 mois (Fig. 3).

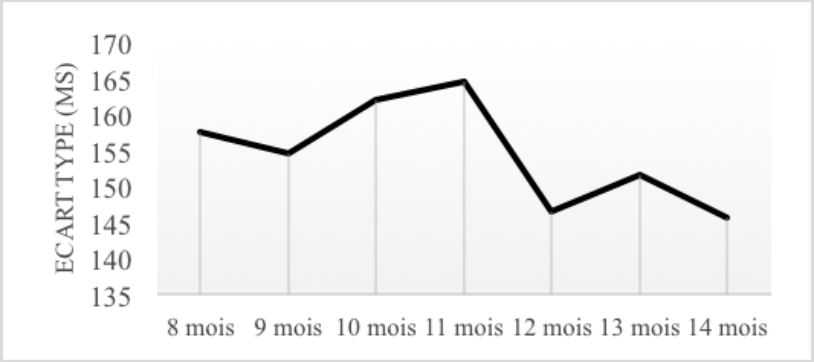



FIGURE 3 : Variation de la durée syllabique (ms) entre l'âge de 8 mois et celui de 14 mois

### 3.3 Comparaison du timing des associations préférentielles et non préférentielles

Lorsque le type d'associations est inclus à l'ANOVA aucune interaction n'est observée entre les variables indépendantes. Les patrons d'évolution temporelle (durée et variabilité) des associations préférentielles et des associations non préférentielles ne sont pas statistiquement différents (durée : $p = 0{,}130$ ; variabilité : $p = 0{,}677$). Néanmoins, certaines tendances relevant de l'analyse qualitative des données méritent d'être pointées (Fig. 4 et 5).

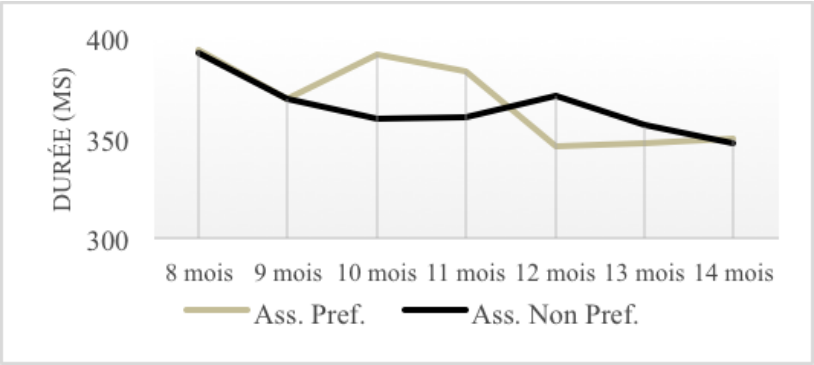



FIGURE 4 : Durée syllabique (ms) des associations préférentielles *vs* non préférentielles entre l'âge de 8 mois et celui de 14 mois

Les syllabes impliquant des associations préférentielles et non préférentielles semblent suivre des profils d'évolution temporelle relativement comparables, mais décalés dans le temps. Ainsi, pour les syllabes constituées de cooccurrences préférentielles, la durée diminue entre 8 et 9 mois pour augmenter à 10 mois et amorcer une nouvelle phase de diminution jusqu'à 14 mois, alors que pour les syllabes constituées de combinaisons non préférentielles, la première phase de diminution s'opère entre 8 et 11 mois, une augmentation intervient à 12 mois à laquelle s'ensuit une nouvelle diminution (Fig. 4). En ce qui concerne la variabilité, les syllabes de type préférentiel voient leur écart type augmenter jusqu'à 11 mois puis chuter jusqu'à 14 mois. Pour les syllabes de type non préférentiel, cette diminution intervient à l'âge de 12 mois (Fig. 5).

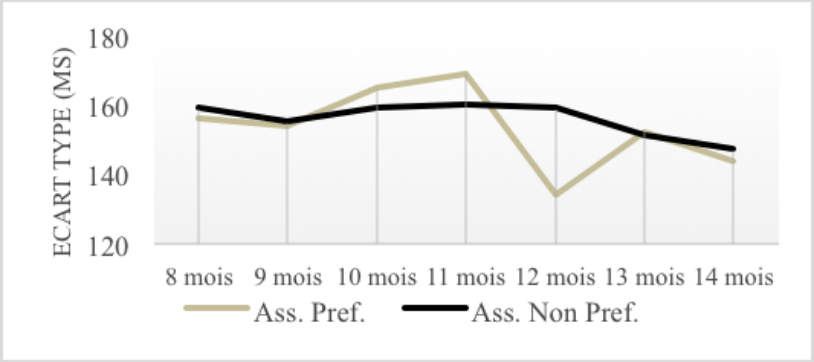



FIGURE 5 : Variation de la durée syllabique (ms) des associations préférentielles *vs* non préférentielles entre l'âge de 8 mois et celui de 14 mois

# 4 Discussion

L'objectif de cette étude était de rendre compte de l'évolution des habiletés oro-motrices au cours de la période du babillage en utilisant la syllabe comme unité d'observation. Plus spécifiquement, nous avons vérifié si la durée et la variabilité temporelle de la syllabe diminuaient avec le temps, et nous avons également évalué l'impact du type d'associations (préférentielles *vs* non préférentielles) sur l'évolution de ces patrons temporels.

La diminution de la durée et de sa variabilité semble confirmer l'augmentation de la rapidité d'exécution de la syllabe et de sa stabilité temporelle entre l'âge de 8 mois et celui de 14 mois (Green et al., 2002 ; Green, Wilson, 2006 ; Nip et al., 2009). Néanmoins, cette diminution n'est pas linéaire (Studdert-Kennedy et al., 1991 ; Smith, Thelen, 2003 ; Green et al., 2010 ; Canault et al., 2011). En effet, à l'âge de 8 mois la durée moyenne de la syllabe est élevée, mais la variation modérée. A 11 mois, la durée poursuit la diminution amorcée à 10 mois, alors que la variabilité temporelle est à son maximum. Puis la durée et la variabilité temporelle de la syllabe diminuent progressivement jusqu'à l'âge de 14 mois. Les résultats obtenus grâce à notre échantillon constitué de plus de 11 000 syllabes, sembleraient confirmer l'existence d'une période critique autour de 10-11 mois (Canault et al., 2011) pouvant correspondre au commencement théorique de la période du babillage varié (Davis, MacNeilage, 2000). Celle-ci pourrait être entrevue comme une phase d'exploration laissant présager l'émergence du contrôle articulatoire pour l'activité de parole (Smith, Thelen, 2003 ; Green et al., 2010). Le processus d'acquisition de la parole répondrait à une succession de phases de progression et de régression impliquant des phases de plus ou moins grande stabilité au cours desquelles une modification des comportements articulatoires est observée. La variation des combinaisons intra syllabiques émergeant à 10 mois et celle des patrons temporels à 11 mois pourraient donc correspondre à une phase d'instabilité qui permettrait au bébé de construire les programmes moteurs des différents patrons syllabiques (Schmidt, 2003). Cette observation, pourrait être imputable au comportement temporel des syllabes impliquant des patrons d'associations préférentielles or celles-ci sont les moins nombreuses à ce stade.

Enfin, même si la durée et la variabilité des syllabes impliquant des patrons associatifs préférentiels et de celles impliquant des patrons associatifs non préférentiels rejoignent des valeurs similaires au terme de la période d'observation, il est intéressant de constater que la diminution de ces paramètres temporels semble s'amorcer plus tôt pour les syllabes qui n'impliquent pas de changement de lieu articulatoire entre la consonne et la voyelle. Cette tendance n'est pas significative, mais reste intéressante à explorer, car on peut s'interroger sur le degré d'influence du développement des

comportements moteurs primaires sur celui des comportements moteurs plus complexes (Rose et al., 2008).

# 5 Conclusion

En dépit de fortes contraintes oro-motrices, la diminution de la durée syllabique et sa stabilisation laissent entrevoir une amélioration du potentiel articulatoire de l'enfant entre l'âge de 8 mois et celui de 14 mois. Ces conclusions étant issues de la généralisation des résultats obtenus chez 22 enfants, il serait intéressant d'analyser les trajectoires individuelles afin de mesurer l'impact de la variabilité inter sujet et de vérifier la diversité ou l'homogénéité des profils développementaux. Notre étude pose néanmoins les bases d'une norme sur le rythme syllabique précoce qui pourrait constituer un paramètre pertinent pour l'évaluation des productions atypiques précoces.

# Remerciements

Les auteurs tiennent à remercier les enfants et leur famille pour leur participation, Marion Hieulle, Sanaé Moinard et Céline Martin pour leur participation au recueil des données et à la segmentation des signaux acoustiques, Jennifer Krzonowski pour l'aide statistique, ainsi que les Labex Aslan et EFL pour leurs soutiens scientifique et financier.

# Références

CANAULT M. LABOISSIÈRE R. (2011). Le babillage et le développement des compétences articulatoires : indices temporels et moteurs. *Faits de Langue,* 37, 173-188.

CANAULT M., LABOISSIÈRE R., PERRIER P., SOCK R. (2008). Development of lingual displacement independence at babbling stage. Actes de *8th International Seminar on Speech Production, ISSP'08*, 177-180.

DAVIS B.L., MACNEILAGE P.F. (1995). The articulatory basis of babbling. *Journal of Speech and Hearing Researc*h, 38, 1199-1211.

DAVIS B.L., MACNEILAGE P.F. (2000). An embodiment perspective on the acquisition of speech perception. *Phonetica*, 57(Special Issue), 229-241.

DOLATA J.K., DAVIS B.L., MACNEILAGE P.F. (2008). Characteristics of the rhythmic organization of vocal babbling: Implications for an amodal linguistic rhythm. *Infant Behavior and Development*, *31*(3), 422–431.

GREEN J.R., MOORE C.A., REILLY K.J. (2002). The sequential development of jaw and lip control for speech. *Journal of Speech, Language, and Hearing Research*, *45*(1), 66–79.

GREEN J.R., WILSON E.M. (2006). Spontaneous facial motility in infancy: A 3D kinematic analysis. *Developmental Psychobiology*, *48*(1), 16–28.

GREEN J.R., NIP I.S., MAASSEN B., VAN LIESHOUT P. (2010). Some organization principles in early speech development. *Speech Motor Control: New Developments in Basic and Applied Research*, 171–188.

KERN S., GAYRAUD F. (2010). *Inventaire Français du Développement Communicatif (IFDC)*. Grenoble : La Cigale.

LIPKIND D., MARCU G.F., BEMIS D.K., SASAHARA K., JACOBY N., TAKAHASI M., SUZUKI K., FEHER O., RAVBAR P., OKANOYA K., TCHERNICHOVSKI O. (2013). Stepwise acquisition of vocal combinatorial capacity in songbirds and human infants. *Nature*, 498, 104–108.

MACNEILAGE P.F. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, *21*(04), 499–511.

MUNHALL K.G., JONES J.A. (1998). Articulatory evidence for syllabic structure. *Behavioral and Brain Sciences*, 21, 524-525.

NIP I.S.B., GREEN J.R., MARX D.B. (2009). Early speech motor development: Cognitive and linguistic considerations. *Journal of Communication Disorders*, 42(4), 286–298.

NIP I.S.B., GREEN J.R., MARX D.B. (2011). The co-emergence of cognition, language, and speech motor control in early development: A longitudinal correlation study. *Journal of Communication Disorders*, *44*(2), 149–160.

PELLEGRINO F., COUPÉ C., MARSICO E. (2011). Across-language perspective on speech information rate. *Language*, *87*(3), 539–558.

ROSE S.A., FELDMAN J.F., JANKOWSKI J.J., ROSSEM R.V. (2008). A cognitive cascade in infancy: Pathways from prematurity to later mental development. *Intelligence*, *36*(4), 367‑378.

SCHMIDT R.A. (2003). Motor Schema Theory after 27 Years: Reflections and implications for a New Theory. *Research Quarterly for Exercise and Sport*, *74*(4), 366–375.

SMITH L.B., THELEN E. (2003). Development as a dynamic system. *Trends in Cognitive Sciences*, *7*(8), 343–348.

STEEVE R.W., MOORE C.A., GREEN J.R., REILLY K.J., MCMURTREY J.R. (2008). Babbling, chewing, and sucking: Oromandibular coordination at 9 months. *Journal of Speech, Language, and Hearing Research*, *51*(6), 1390–1404.

STUDDERT-KENNEDY M., KRASNEGOR D.M., RUMBAUGH R., SCHEIFELBUSCH R. (1991). Language development from an evolutionary perspective. *Biological and Behavioral Determinants of Language Development*, 5–28.

VIHMAN M.M. (1992). Early syllables and the construction of phonology. In Charles Ferguson, Lise Menn and Carol Stoel-Gammon (Eds.), *Phonological Development: Models, Research, Implications*. Timonium, MD: York Press.

YAMAGUCHI N., DOS SANTOS C., KERN S. (2015). Ce que révèle l'ordre d'acquisition des classes naturelles à propos des harmonies consonantiques. *Lidil. Revue de Linguistique et de Didactique des Langues*, 51, 89-117.

# Étude exploratoire des événements articulatoires pendant la réalisation de pauses en parole spontanée

Ivana Didirková[1], Camille Fauth[2], Sébastien Le Maguer[3]
(1) F.R.S.-FNRS & VALIBEL, Université Catholique de Louvain, Belgique
(2) Université de Strasbourg, Institut de Phonétique, E.A. 1339 LiLPa, Strasbourg, France
(3) Saarland University, Allemagne
ivana.didirkova@gmail.com

## RÉSUMÉ

Les disfluences en tant qu'événements arrêtant momentanément le déroulé du discours font partie intégrante de la production de la parole. De nombreuses études ont analysé leurs caractéristiques en se fondant principalement sur leurs propriétés acoustiques. L'objectif de cette étude est d'enrichir la description acoustique des disfluences produites par quatre locuteurs en parole spontanée en la corrélant à une description articulatoire, obtenue à partir d'un articulographe électromagnétique (EMA), synchronisée avec les données acoustiques. Pour ce faire, nous avons observé leur production à l'aide de deux différentes méthodes : une étude articulatoire effectuée de façon automatique à partir de la vitesse des mouvements des articulateurs et une observation experte du mouvement de ces mêmes articulateurs. Les résultats de cette étude exploratoire montrent l'existence de plusieurs configurations articulatoires pour un même type perçu de disfluence et inversement, plusieurs catégories acoustiques peuvent parfois être traduites par une même configuration articulatoire. De même, l'anticipation du geste articulatoire relatif au phone subséquent semble dépendre de l'activité articulatoire supra-glottique.

## ABSTRACT

**An exploratory study of articulatory events during pause realization in spontaneous speech.**
Disfluencies, as a momentary stop event in a discourse, are considered as a constitutive part of speech production. Numerous studies analyzed their specificities by relying mainly on their acoustic properties. The goal of the presented study is to extend the acoustic analysis by including a articulatory description. This description is obtained using Electro-Magnetic Articulography (EMA) synchronized with the acoustic data. To achieve this study, we have observed the production of the disfluencies following two complementary methodologies: an automatic articulatory study based on the quantification of the velocity of the articulators and an expert subjective analysis of the movement of these articulators. The results of this exploratory study show the same perceived disfluency category can rely on multiple articulatory configurations. Multiple acoustic categories can also be realized following a same articulatory configuration. Furthermore, the anticipation of the subsequent gesture seems to be dependent on the supraglottic activity.

MOTS-CLES : Production de la parole, pauses, disfluences, description articulatoire, EMA
KEYWORDS: Speech production, pauses, disfluencies, articulatory description, EMA


10.21437/JEP.2018-51

# 1 Introduction

La communication ne se fonde pas exclusivement sur les mots employés ou sur les éléments prosodiques retenus par le locuteur. D'autres événements paralinguistiques, plus ou moins intentionnellement produits, participent à la construction du discours. Parmi ces éléments, notre attention se porte ici sur les disfluences (ou accidents de parole) qui viennent interrompre le discours et peuvent donc parfois entraver son intelligibilité (Mac Gregor *et al*., 2009). Traditionnellement, deux types de disfluences sont observées (Fromkin & Ratner, 1998) : les pauses silencieuses (ou vides), qui sont une interruption du flux de parole se répercutant sur le signal acoustique par une amplitude nulle ou non-significative (Duez, 2003), et les pauses remplies ou pleines qui regroupent les répétitions, les faux-départs, les allongements de sons indépendants de l'allongement final de fin de syntagme (Duez, 2001), la réalisation d'un *schwa* (Maclay & Osgood, 1959 par ex.) ou l'utilisation d'un autre type de filler. De nombreuses études ont cherché à quantifier et à analyser ces disfluences, montrant que celles-ci sont d'autant plus fréquentes que le discours n'est pas préparé à l'avance (Corley & Stewart, 2008). De plus, l'endroit où se réalisent les disfluences ne serait pas anodin dans la mesure où elles ont plus de chance d'apparaître en début de prise de parole ou en début de phrase (Maclay & Osgood, 1959 ; Beattie & Bradbury, 1979) ou lorsque les énoncés à venir sont longs (Oviatt, 1995 ; Shriberg, 1996) ou enfin que le locuteur est peu familier du sujet discuté (Bortfeld *et al*., 2001 ; Merlo & Mansur, 2004).

Toutefois, la plupart de ces études se fondent sur des données acoustiques et peu de recherches se sont portées sur les évènements articulatoires pendant les disfluences. Signalons toutefois l'étude de Ramanarayanan *et al*., (2009) qui ont observé le tractus vocal pendant la réalisation de pauses silencieuses grâce à l'IRM dynamique et montré que le comportement des articulateurs diffère en fonction de la nature de la pause. Ainsi, la vitesse des articulateurs décroit significativement lorsque la pause est planifiée alors que ce n'est pas le cas pour les pauses agrammaticales. Ces résultats sont à rapprocher de l'étude pilote de Lalain *et al*. (2016) qui a montré que les données articulatoires semblent pouvoir également contribuer à l'identification d'indices robustes pour caractériser les pauses respiratoires (abaissement systématique de la mandibule) et la déglutition (élévation de l'apex, du dos et du larynx).

Ainsi, ces travaux, généralement conduits à partir de données acoustico-perceptives, concluent qu'il existerait deux types de pauses : les pauses vides (qui peuvent-être syntaxiques ou non) et les pauses remplies. Notre *objectif*, avec cette étude, est d'observer si la dichotomie, fondée sur des indices acoustiques et perceptifs, entre pauses vides et pauses pleines correspond à l'activité articulatoire supra-glottique. En d'autres termes, est-ce que des patterns articulatoires spécifiques aux différents types pauses peuvent être observés ? Pour cela, nous nous baserons sur deux critères : présence / absence de mouvements articulatoires et anticipation de mouvement relatif au phone subséquent à la disfluence. Notre *hypothèse* est que, à l'instar des résultats présentés dans Didirková (2016) sur des données portant sur la production de la parole par des sujets qui bégaient, les catégories acoustico-perceptifs ne devraient pas être représentatives de l'activité articulatoire.

# 2 Méthodologie

## 2.1 Corpus et participants

Notre étude s'appuie sur des enregistrements articulatoires acquis à l'aide d'un articulographe électromagnétique de type Carstens AG501 3D au LORIA avec une fréquence d'échantillonnage de

250 Hz et une précision constructeur de 0,3 mm. Ces données ont été stockées sous format *.pos* et synchronisées avec des données acoustiques (format *.wav*, 44,1 kHz, 16 bits) obtenues à l'aide d'un microphone t.bone EM 9600. Les enregistrements se sont déroulés dans un endroit calme.

Quatre sujets ont pris part à cette étude, 2 femmes et 2 hommes de langue maternelle française, appariés en catégorie socio-professionnelle et en âge (âge moyen : 31,25 ans, ET : 5,12). Après la fixation des capteurs (1 au milieu de chaque lèvre, 1 sur la mandibule, 1 sur la pointe de la langue, 1 sur le pré-dos et 1 sur le dos de la langue, plus 3 capteurs servant à contrôler les mouvements de la tête dont 1 sur le front et 1 derrière chaque pavillon auriculaire), plusieurs tâches de production leur ont été demandées. Les analyses présentées dans le cadre de cette étude portent sur la parole spontanée (description d'une journée type, des passe-temps avec relance de l'expérimentateur si nécessaire) et ce, dans l'objectif d'obtenir des disfluences liées à la programmation linguistique. Les données ainsi acquises constituent un corpus d'une durée totale de 26 min 02 s, allant de 5 min 5 s à 8 min 14 s selon le sujet. À l'intérieur de ces laps de temps, les tours de parole des sujets (après exclusion des interventions de l'expérimentateur) représentent 21 min et 19 s au total.

## 2.2 Traitement de données

L'ensemble du corpus a été transcrit orthographiquement puis segmenté et annoté de manière semi-automatique à l'aide du logiciel Praat (Boersma & Weenink, 2017) et de l'outil d'alignement automatique et phonétique EasyAlign (Goldman, 2011). Une ligne d'annotation spécifique a été ajoutée pour y repérer les différents types de disfluences. Comme le remarque Candea (2000), ces éléments peuvent renvoyer, selon les auteurs, à des phénomènes différents. Pour ce travail, nous avons considéré les pauses selon les catégories suivantes : les prolongations correspondant aux allongements audibles d'un son (y compris celles des marqueurs d'hésitation de type « euh »), les répétitions de phonèmes et les pauses silencieuses (vides) syntaxiques ou non-syntaxiques (disfluentes). Afin de faciliter la lecture, toutes ces catégories vont être regroupées sous le terme de « disfluences », même si l'inclusion des pauses syntaxiques dans ce groupe peut faire débat. Ces catégories ont été choisies car, ne concernant que l'absence d'articulation ou un seul phone, elles permettent de minimiser l'influence de la coarticulation sur les analyses articulatoires. Les pauses vides ont été annotées sans seuil minimum.

Les analyses reposent ainsi sur une identification acoustico-perceptive des disfluences susmentionnées par deux auteurs. Les cas de désaccord ont été discutés. Les sujets ont produit 714 disfluences (voir TABLEAU 1 pour la distribution). Les pauses vides contenant des bruits de bouche (au nombre de 78) ont été exclues de la suite des analyses car possiblement provoqués par les capteurs collés dans la cavité buccale. Les analyses qui suivent couvrent donc 636 disfluences.

| Locuteur | PV | SD | P | REPph | **Total** |
|---|---|---|---|---|---|
| F1 | 46 | 16 | 89 | 0 | 151 |
| F2 | 33 | 17 | 57 | 0 | 107 |
| M1 | 57 | 25 | 141 | 6 | 229 |
| M2 | 95 | 13 | 40 | 1 | 149 |
| **Total** | 231 | 71 | 327 | 7 | **636** |

TABLEAU 1: Distribution des disfluences par locuteur et par type de disfluence. F - locuteur féminin, M - locuteur masculin. PV - pauses vides syntaxiques, SD - pauses vides non-syntaxiques, P - prolongations, REPph - répétitions de phonèmes.

Les TextGrids ont par la suite été exportés dans le logiciel Visartico (Ouni *et al*., 2012) permettant une visualisation des données articulatoires dans différentes configurations, dont une vue en 2D avec coupe midsagittale ou en 3D. Ce logiciel rend également possible l'obtention d'un suivi de l'évolution spatio-temporelle des articulateurs synchronisé avec le signal de parole. Précisons que seules les *tiers* phone, words et disfluences ont été visualisées dans Visartico (FIGURE 1).

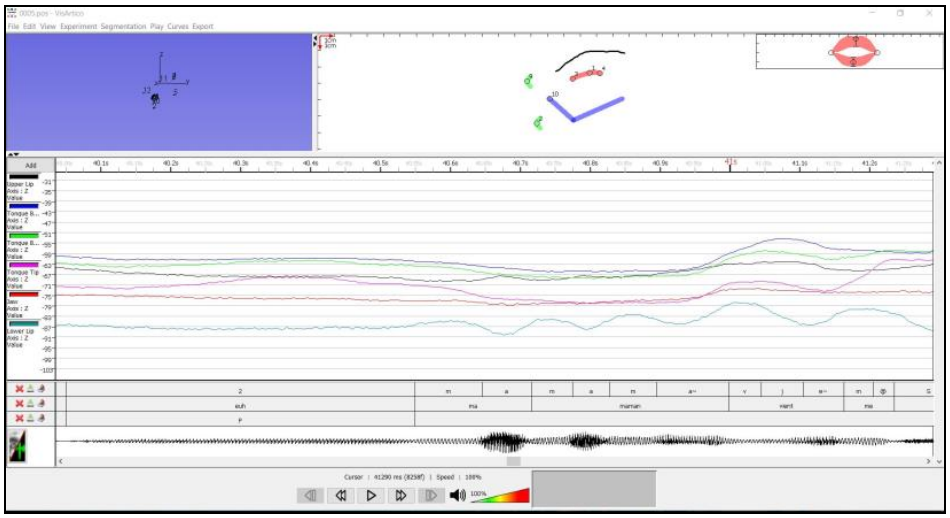


FIGURE 1 : Exemple d'annotation visualisé dans Visartico. En haut à gauche, une vue 3D des mouvements des capteurs. À droite, une vue sagittale médiane ainsi qu'une vue frontale des lèvres. Au centre, l'évolution spatio-temporelle des mouvements articulatoires verticaux. En bleu, les mouvements du dos de la langue ; en vert, les mouvements du prédos ; en noir, la lèvre supérieure ; en rose, l'apex ; en rouge, la mandibule et en gris, la lèvre inférieure.

Après l'importation des données, le logiciel Visartico a servi à la visualisation des mouvements articulatoires verticaux des articulateurs supra-glottiques mentionnés *supra* durant les disfluences. Ces mouvements ont par la suite été décrits puis classés dans différentes catégories. Enfin, l'extension de l'anticipation du mouvement lié au phone suivant immédiatement la disfluence a été calculée. Pour cette dernière mesure, les analyses portent systématiquement sur le mouvement de l'articulateur principal nécessaire à la production du phone suivant.

Signalons que l'analyse des signaux EMA a motivé l'exclusion de 12 occurrences de disfluences dans l'étude portant sur les *patterns* observés et ce, pour cause des anomalies sur le signal. En outre, l'anticipation n'a pas pu être analysée pour 90 disfluences puisqu'elles étaient suivies d'une pause silencieuse ou d'une fin de tour de parole.

## 2.3 Analyse automatique

Pour déterminer si une trame est en mouvement, nous avons choisi un critère basé sur la vitesse instantanée, détaillé dans Didirková *et al*. (2017). Tout d'abord, nous assimilons la vitesse instantanée à la dérivée première qui est calculée de manière discrète, telle qu'implémentée dans le système HTK[1]. Ensuite, une trame est considérée comme étant en mouvement si sa dérivée dépasse un certain seuil. Ce seuil a été arbitrairement défini à 30% de la vitesse moyenne calculée sur l'ensemble des trames associées au locuteur analysé et ce, afin d'encourager la détection de

[1] Voir http://htk.eng.cam.ac.uk pour plus de détails.

mouvement. Cette étape effectuée, l'analyse est poursuivie à l'échelle du segment. L'objectif de notre analyse est de vérifier si un mouvement se produit dans un segment supposé statique. Ainsi, afin d'éviter les détections dues aux transitions, les premières trames ainsi que les dernières trames du segment concerné (soit 10% au début et à la fin du segment) ont été écartées.
L'analyse automatique est fondée sur deux mesures. La première consiste à calculer le ratio de trames en mouvement par rapport au nombre total de trames du segment. La seconde se focalise sur l'amorce du premier mouvement significatif. Pour cela, nous avons arbitrairement défini que si 10 trames (soit 50 ms) consécutives étaient en mouvement, alors nous étions en présence d'un mouvement significatif. Ainsi, nous pouvons calculer à quelle proportion du segment le premier mouvement débute.

# 3 Résultats

## 3.1 Pourcentage de trames en mouvement en fonction du type perceptif des disfluences

La première analyse automatique porte sur le pourcentage de disfluences considérées comme étant en mouvement (FIGURE 2). Cette première étude montre que le ratio de disfluences caractérisées par la présence d'un mouvement articulatoire supra-glottique atteint les 80% dans tous les types perceptifs des disfluences.

L'analyse a ensuite été affinée en s'intéressant au nombre de trames en mouvement par type de disfluence. Cette analyse plus détaillée est présentée dans la FIGURE 3. Même si globalement une grande majorité de disfluences présentent des mouvements, cela ne signifie pas pour autant que toutes les trames sont en mouvement suivant les critères définis *supra*. En effet, une tendance des silences non-syntaxiques à avoir davantage de trames en mouvement (80% en moyenne) que les pauses vides syntaxiques ou les prolongations peut être observée.

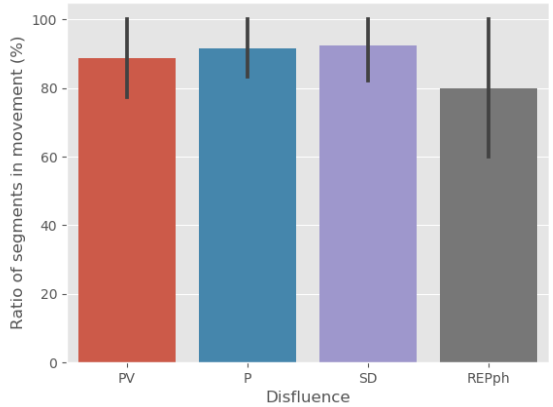



FIGURE 2 : Pourcentage de disfluences considérées comme étant en mouvement.

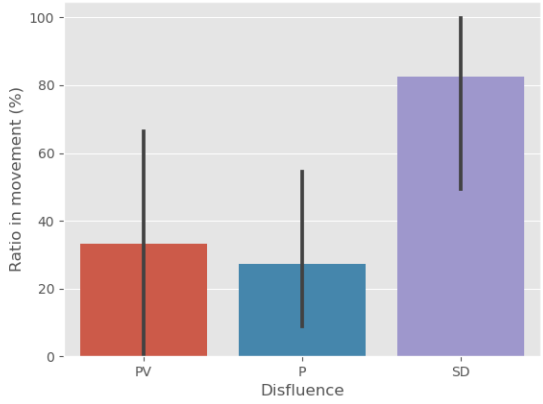



FIGURE 3 : Pourcentage de trames considérées comme étant en mouvement en fonction du type perceptif de la disfluence.

## 3.2 Patrons articulatoires

L'analyse experte manuelle des mouvements verticaux présents dans les disfluences a permis de faire ressortir l'existence de trois grands patrons articulatoires (test de chi2 goodness-of-fit, $p < 0,01$). En effet, les disfluences analysées peuvent se traduire par le maintien global de la posture articulatoire avec ou sans sortie acoustique, la présence desdits mouvements tout au long de la disfluence, ou une combinaison des deux premiers patrons et ce, dans un ordre pouvant varier.
Cela étant, il est à signaler que les sujets manifestent une nette préférence pour le pattern où des mouvements articulatoires sont présents tout au long de la disfluence (89,94% de toutes les disfluences) par rapport à un maintien global de la posture articulatoire (6,16%) et à une combinaison des deux premiers patterns (3,9%). Cette observation confirme par ailleurs celle qui a été faite à partir de la détection automatique de mouvements, présentée *supra*.
Lorsque l'on s'intéresse à la distribution des patrons articulatoires susmentionnés par rapport au type perceptif de la disfluence, on observe sans surprise que le patron caractérisé par la présence de mouvements peut être trouvé dans les quatre types analysés (FIGURE 4). En ce qui concerne le maintien global de la posture articulatoire, ce dernier peut principalement être observé dans les prolongations, alors que la combinaison des deux était surtout présente dans les pauses non-syntaxiques. Conformément au test de chi2 pour l'indépendance, cette corrélation est significative ($p < 0,01$) mais l'effet n'est pas particulièrement fort (0,256).
L'analyse des durées des disfluences permet de montrer également une tendance du pattern « combinaison » à apparaître lorsque la disfluence est plus longue (FIGURE 5). Ceci est vrai pour tous les types perceptifs des disfluences. Inversement, lorsque la durée de la disfluence diminue, le comportement des types perceptifs des disfluences n'est pas le même. En effet, l'on observe notamment que les prolongations ainsi que les pauses vides non-syntaxiques les plus courtes sont caractérisées par la présence de mouvements articulatoires alors que dans les pauses vides syntaxiques les plus courtes, c'est un maintien global de la posture articulatoire qui est le plus représenté dans notre corpus.

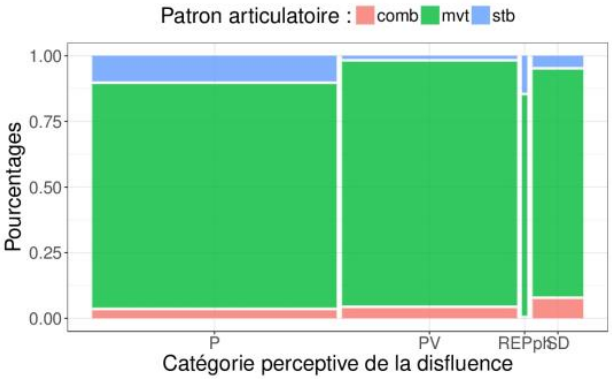



FIGURE 4 : Distribution des patrons articulatoires en fonction du type perceptif de la disfluence. Comb = combinaison (en rouge), mvt = mouvements (en vert), stb = maintien global de la posture (en bleu). Sur l'ordonnée, la proportion observée (de 0 à 1). Sur l'abscisse, le type perceptif de la disfluence.

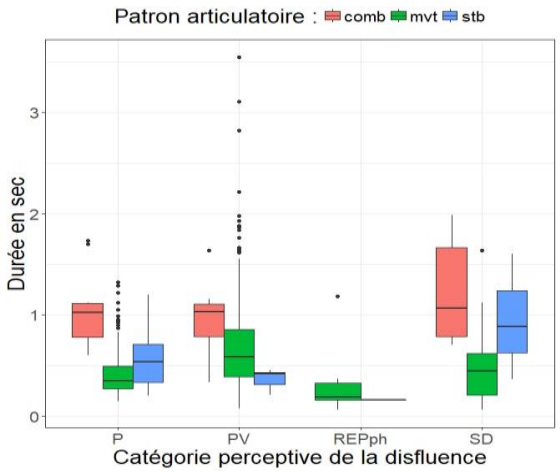



FIGURE 5 : Distribution des patrons articulatoires en fonction de la durée de la disfluence. Comb = combinaison (en rouge), mvt = mouvements (en vert), stb = maintien global de la posture (en bleu). Sur l'ordonnée, la durée de la disfluence (en secondes). Sur l'abscisse, le type perceptif de la disfluence.

### 3.3 Anticipation dans les disfluences

La dernière analyse présentée dans cette étude concerne l'extension de l'anticipation dans les disfluences. Rappelons que l'objectif de cette analyse était de s'intéresser à la manière dont les phones suivant immédiatement la disfluence sont anticipés du point de vue temporel. Les résultats, présentés dans la FIGURE 6, montrent que l'extension de l'anticipation (en % sur la durée totale de la disfluence) est plus importante lorsque les articulateurs ont été en mouvement tout au long de la disfluence et ce, pour les prolongations et les pauses vides non-syntaxiques, laissant penser à une sorte de continuité dans les mouvements. Pour ces deux catégories perceptives, soulignons également que le maintien global de la position des articulateurs était le patron articulatoire qui se prêtait le moins à une anticipation précoce. Dans ces cas-là, le démarrage du geste anticipatoire se fait à moins de 25% de la disfluence. En revanche, dans les pauses vides syntaxiques, la tendance semble inversée dans la mesure où c'est en situation de maintien de posture que l'anticipation vient le plus tôt. Cette différence pourrait s'expliquer par le fait que dans les pauses vides syntaxiques, le degré de préparation du message linguistique qui suit est plus important que dans d'autres types de disfluences. Toutefois, la variation est très importante dans tous les cas de figure.
Signalons que 97 disfluences ont été exclues de cette analyse car suivies d'une pause vide.

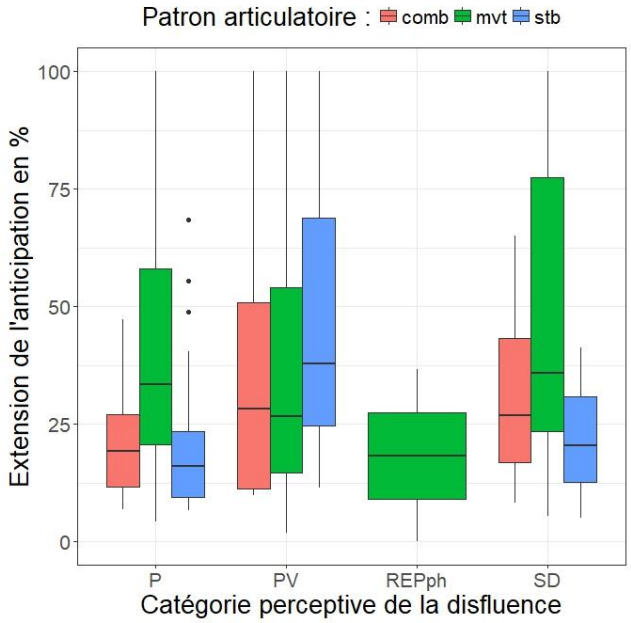



FIGURE 6: Extension de l'anticipation en fonction du type perceptif et du patron articulatoire. Comb = combinaison (en rouge), mvt = mouvements (en vert), stb = maintien global (en bleu). Sur l'ordonnée, l'extension de l'anticipation (en %). Sur l'abscisse, le type perceptif de la disfluence.

# 4 Discussion et conclusion

Rappelons que l'objectif de cette étude était double. Il s'agissait (1) de vérifier si la catégorisation des disfluences en fonction de critères acoustiques et perceptifs correspond à l'activité articulatoire supra-glottique et (2) de compléter la description de ces événements de la parole du point de vue de l'extension de l'anticipation du geste liée à la production du phone immédiatement subséquent à la disfluence.

Dans la première partie de l'étude, deux méthodes complémentaires ont été utilisées ; l'une basée sur des critères reposant sur la vitesse et permettant une analyse automatique, l'autre fondée sur une analyse manuelle des courbes de l'évolution spatio-temporelle des articulateurs supra-glottiques. Les deux méthodes ont permis de montrer que les disfluences présentes en parole spontanée se caractérisent, pour la plupart, par la présence d'une activité articulatoire supra-glottique. L'analyse manuelle a montré en outre que cette activité peut parfois être couplée avec une phase de maintien de la posture articulatoire au sein d'une même disfluence. Par ailleurs, un certain nombre de disfluences ne comportaient pas de mouvement détectable automatiquement ou manuellement. De fait, il peut être avancé que la dénomination des disfluences fondée sur les critères acoustiques et perceptifs ne reflète pas suffisamment la réalité articulatoire supra-glottique. À titre d'exemple, les prolongations ne sont pas systématiquement de simples « maintiens » de la posture articulatoire mais se caractérisent par la présence de mouvements articulatoires. De l'autre côté, les pauses ne sont pas exemptes de toute activité articulatoire, comme l'ont déjà souligné Lalain *et al*. (2016). Enfin, si les répétitions de phones n'étaient pas nombreuses dans le corpus, elles ont laissé voir que la répétition n'est pas toujours une réitération du mouvement articulatoire mais que parfois, les articulateurs peuvent être dans une position globalement stable. Dans ce cas, on peut supposer que la répétition sera due davantage à un acte phonatoire qu'articulatoire.

L'étude de l'extension de l'anticipation a également permis d'observer quelques tendances, montrant notamment que le patron articulatoire aurait une influence sur le degré d'anticipation dans la mesure où celui-ci était plus important lorsqu'une activité articulatoire était présente dans la disfluence. Cette tendance n'a cependant pas été confirmée pour les pauses syntaxiques, montrant que la préparation du message linguistique est un facteur à inclure dans les analyses à venir.

# 5 Perspectives

Si cette étude a permis d'observer certains phénomènes liés à la production des événements arrêtant momentanément le déroulé du discours, il est important de souligner que le nombre de locuteurs doit être augmenté afin de confirmer les tendances observées et éventuellement les compléter par d'autres patterns. En outre, il serait intéressant d'élargir l'analyse à d'autres critères que les mouvements verticaux utilisés pour les analyses manuelles et la vélocité employée pour celles automatiques. Par ailleurs, une analyse portant sur la vélocité avant et après les disfluences apporterait davantage de précisions sur la manière dont les planifications phonologique et articulatoire est faite en phase de disfluence en parole spontanée. L'étude portant sur le nombre de trames en mouvement détecté de manière automatique a aussi fait ressortir l'importance d'une analyse plus fine pour les disfluences considérées comme étant en mouvement. Nous avons notamment pu observer une tendance pour les silences non-syntaxiques à comporter un plus grand nombre de trames en mouvement que les prolongations et les pauses vides syntaxiques. Il semble ainsi intéressant d'affiner le pattern « mouvement » en fonction de plusieurs critères, tels que nombre d'articulateurs en mouvement ou encore la distinction entre l'articulateur principal (qui contribue directement à la production du phone en cours) et les autres. Enfin, les deux méthodes, automatique et manuelle, ayant démontré leur complémentarité, il serait profitable de développer, à terme, une analyse automatique permettant de guider l'analyse experte.

# Références


BEATTIE G. W., BRADBURY R. J. (1979). An experimental investigation of the modifiability of the temporal structure of spontaneous speech. *Journal of Psycholinguistic Research*, *8*(3), 225-248.

BORTFELD H., LEON S. D., BLOOM J. E., SCHOBER M. F., BRENNAN S. E. (2001). Disfluency rates in conversation: effects of age, relationship, topic, role, and gender. *Language and Speech*, *44*(Pt 2), 123-147.

CORLEY M., STEWART O W. (2008). Hesitation Disfluencies in Spontaneous Speech: The Meaning of um. *Language and Linguistics Compass*, *2*(4), 589-602.

CANDEA M. (2000). *Contribution à l'étude des pauses silencieuses et des phénomènes dits d'"hésitation" en français oral spontané*, Thèse de Doctorat, Paris 3 Sorbonne Nouvelle.

DIDIRKOVÁ I. (2016). *Parole, langues et disfluences : une étude linguistique et phonétique du bégaiement*, Thèse de Doctorat, Université Paul-Valéry Montpellier 3.

DIDIRKOVÁ I., LE MAGUER S., GBEDAHOU, D., HIRSCH F. (2017). What happens during stuttering-like disfluencies? An EMA study. *International Seminar on Speech production*, Tianjin, Chine.

DUEZ D. (2001). Caractéristiques acoustiques et phonétiques des pauses remplies dans la conversation en français. *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA)*, *20*, 31-48.

DUEZ D. (2003). Le pouvoir du silence et le silence du pouvoir : comment interpréter le discours politique. *MediaMorphoses*, (8), 77-82.

FROMKIN V. A., RATNER N. B. (1998). Speech production. In J. B. Gleason & N. B. Ratner (Éd.), *Psycholinguistics*. Harcourt Brace College Publishers.

LALAIN M., LEGOU T., FAUTH C., HIRSCH F., DIDIRKOVÁ I. (2016). Que disent nos silences ? Apport des données acoustiques, articulatoires et physiologiques pour l'étude des pauses silencieuses. *JEP TALN RECITAL 2016*. Paris, France.

MAC GREGOR L., CORLEY M., DONALDSON D. I. (2009). Not all disfluencies are are equal: The effects of disfluent repetitions on language comprehension. *Brain and Language*, *1*(111), 36-45.

MACLAY H., OSGOOD C. E. (1959). Hesitation Phenomena in Spontaneous English Speech. *WORD*, *15*(1), 19-44.

MERLO S., MANSUR L L. (2004). Descriptive discourse: topic familiarity and disfluencies. *Journal of Communication Disorders*, *37*(6), 489-503.

OUNI, S., MANGEONJEAN L., STEINER I. (2012). VisArtico: a visualization tool for articulatory data, *Interspeech2012*, September 9-13, 2012, Portland, OR, USA.

OVIATT S. (1995). Predicting and Managing Spoken Disfluencies During Human-Computer Interaction. *Computer Speech & Language*, *9*(1), 19-35.

RAMANARAYANAN V., BRESCH E., BYRD D., GOLDSTEIN L., NARAYANAN S. S. (2009). Analysis of pausing behavior in spontaneous speech using real-time magnetic resonance imaging of articulation. *The Journal of the Acoustical Society of America*, *126*(5).

SHRIBERG E. (1996). Disfluencies In Switchboard. *Proceedings International Conference on Spoken Langage Processing* (Vol. Addendum, p. 11-14). Philadelphie.

# La parole sans les lèvres : une étude acoustique et articulatoire

Hannah King[1] Emmanuel Ferragne[1]
(1) Université Paris Diderot, CLILLAC-ARP, 5 rue Thomas Mann, 75013 Paris, France
hannah.king@univ-paris-diderot.fr, emmanuel.ferragne@univ-paris-diderot.fr

## RESUME

Cet article évalue le potentiel d'un écarteur de lèvres pour les études phonétiques sur la perturbation. Cet appareil est désormais utilisé dans le « défi de l'écarteur de bouche » qui est devenu un véritable phénomène internet. Lors du port de l'appareil, les mouvements des lèvres sont rendus impossibles. Les données acoustiques et articulatoires de quatre locuteurs d'anglais britannique sont présentées. Un accéléromètre est utilisé pour évaluer la dynamique de la mâchoire et l'échographie linguale nous donne les informations sur les éventuelles stratégies de compensation, notamment pour la voyelle /u/. Grâce aux données échographiques, nous observons une rétraction de la langue pour /u/ perturbé, ce qui ne figure pas dans les données acoustiques correspondantes. Cette étude souligne donc les limitations d'une analyse purement acoustique de la parole perturbée. Malgré ses inconvénients, nous concluons que l'utilisation d'un écarteur de lèvres est une technique prometteuse pour la recherche sur la perturbation labiale.

## ABSTRACT

**Speech without lips: an acoustic and articulatory study**
This paper studies the use of a lip retractor as a potential technique for phonetic studies involving perturbation. This device is currently used by participants of the internet sensation, the so-called "no lips" or "mouth guard" challenge. Wearing the device restricts the use of the lips during speech. We present acoustic and articulatory data from four speakers of British English. Accelerometer data is used to assess the dynamics of the jaw and ultrasound tongue imaging gives us insights into potential compensation strategies, specifically for the /u/ (GOOSE) vowel. Ultrasound data revealed that three speakers showed signs of tongue retraction for perturbed /u/, which was not reflected in the corresponding acoustic data. This study highlights the limitations of a purely acoustic analysis of the effects of perturbation on speech. Despite certain limitations, we conclude that the use of the lip retractor is a promising technique for future lip perturbation studies.

MOTS-CLES : Perturbation labiale, stratégies de compensation, échographie linguale, corrélats acoustiques-articulatoires, anglais britannique, antériorisation de GOOSE

KEYWORDS: Lip perturbation, compensation strategies, ultrasound tongue imaging, acoustic-articulatory correlates, British English, GOOSE fronting

# 1 Introduction

De nombreuses études se sont intéressées à la perturbation mécanique des articulateurs afin de répondre à certaines questions importantes sur la nature de la production de la parole. Une perturbation dans la cavité orale nous oblige à développer de nouvelles stratégies pour produire le son désiré. Les études précédentes ont démontré que les locuteurs sont capables de reproduire une


10.21437/JEP.2018-52

cible acoustique donnée en ayant recours à une réorganisation des articulateurs, ce qui suggère que la production est de nature auditive et non pas articulatoire. Par exemple, Riordan (1977) a observé un geste laryngal atypique suite à une perturbation des lèvres pour /y/ en français. Cependant le degré de succès de la compensation peut varier. Dans une autre étude avec une perturbation des lèvres, malgré des tentatives de reconfiguration des articulateurs, certains participants ont finalement privilégié leur articulation canonique et par conséquent, n'ont pas réussi à optimiser leur production de /u/ perturbé (Savariaux et al., 1995 ; Savariaux et al., 1999). Nos stratégies articulatoires habituelles jouent donc un rôle dans les primitives phonologiques de la production de la parole (Perrier, 2005).

Plusieurs dispositifs ont été employés pour créer une perturbation mécanique dans le conduit vocal, y compris des cale-dents, des tubes tenus entre les lèvres, et des prothèses palatales. Cet article s'intéresse à l'effet d'un écarteur de lèvres (également connu sous le nom d'embout ou d'ouvre bouche), un appareil qui se place entre les lèvres et les dents de manière à ce qu'il touche l'intérieur de la joue. Il ouvre les lèvres latéralement, limitant donc les mouvements des lèvres et de la mâchoire. Bien qu'il ait été initialement prévu pour l'orthodontie, cet appareil est actuellement porté par les participants du « défi de l'écarteur de bouche ». Il s'agit d'un véritable phénomène internet sur lequel la marque Hasbro a récemment créé un jeu de société appelé « Mâche Mots ». On y tente de faire deviner aux autres joueurs des mots et expressions produits en portant ce dispositif[1].

A notre connaissance, aucune étude n'a été menée sur l'effet de cet appareil sur la production de la parole. Nous cherchons à évaluer le potentiel de cet appareil pour les études futures sur la perturbation. Nous présenterons une analyse articulatoire et acoustique préliminaire de l'effet de l'écarteur de lèvres sur les monophtongues de l'anglais britannique chez quatre locuteurs. Pour recueillir les données articulatoires, nous avons utilisé l'échographie de la langue et un accéléromètre pour analyser les mouvements de la mâchoire. Nous examinerons enfin si les locuteurs ont réussi à opérer une compensation pour la perturbation de la voyelle /u/, qui nécessite habituellement le recours à un certain degré de labialisation.

# 2 Effet de la perturbation

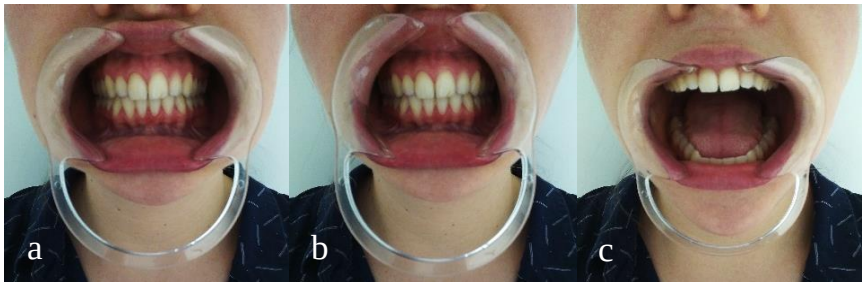



FIGURE 1 : L'écarteur de lèvres : a) position neutre b) arrondissement c) bouche ouverte

Dans la FIGURE 1, l'image a) présente la bouche dans une position neutre (bouche fermée, sans arrondissement des lèvres). On remarque que l'appareil ajoute des canaux latéraux : un espace important entre l'intérieur des joues et les dents. L'écarteur a une certaine souplesse et la taille de ces canaux latéraux semble diminuer lorsqu'on essaie d'arrondir les lèvres (image b de la FIGURE 1).

[1] Pour un exemple du jeu : https://youtu.be/pN866ZoYqS4

Cette différence peut s'expliquer par le fait que l'intérieur des commissures des lèvres restent normalement en contact avec la surface des dents pendant un arrondissement des lèvres (Li et al., 2015). Ce qu'on observe en b) serait donc un mouvement de rapprochement des commissures. Comme le montre la FIGURE 1, cet appareil rend un arrondissement, un étirement et une constriction des lèvres impossibles. En revanche, l'appareil n'entrave pas l'ouverture de la bouche. L'image c) montre son ouverture maximale lors du port de l'appareil.

Des études précédentes ont simulé l'effet de la perturbation sur la parole à partir de modélisations acoustiques du conduit vocal (Savariaux et al., 1995). Ces modélisations ont également été employées pour prédire les stratégies de compensation. Dans cette approche, l'obtention de données géométriques réelles du conduit vocal est essentielle (Ghio A., 2007). Ne disposant pas de telles données, nous ne pouvons pas adopter une telle approche pour modéliser l'effet de l'écarteur. Vu l'ajout des canaux latéraux qui élargissent le conduit vocal, nous ne pouvons pas simplement retirer les lèvres des modélisations acoustiques existantes, comme celles de Fant (1960) et Stevens (1989). Par conséquent, nos prédictions concernant l'impact de l'écarteur sur les monophtongues de l'anglais ne peuvent être que générales. Premièrement, vu que l'appareil empêche l'arrondissement des lèvres, les voyelles arrondies ne devraient pas comporter d'arrondissement. Ensuite, bien que l'appareil étire les lèvres, il semble ne pas permettre la production de voyelles très écartées, comme [i]. De plus, l'ajout des canaux latéraux pourrait entrainer de la friction. Enfin, même si l'appareil présente une certaine souplesse, nous prévoyons également une perturbation de la cinématique mandibulaire, que nous allons quantifier au moyen d'un accéléromètre.

Afin de juger si les locuteurs arrivent à opérer une compensation de la perturbation, nous nous concentrons sur la voyelle /u/. Cette voyelle est en effet une variable sociolinguistique importante dans plusieurs accents anglais. Elle a suscité de nombreuses études car elle est souvent antériorisée par rapport à sa catégorisation phonologique (/u/, donc conventionnellement fermée et postérieure) et par rapport à sa réalisation phonétique historique (Scobbie et al., 2012). Les études qui traitent de l'antériorisation de /u/ (également connue sous le terme « GOOSE fronting ») s'appuient généralement sur les données acoustiques, acceptant donc le parallèle entre trapèze vocalique articulatoire et plan F1/F2. Cependant on sait que les caractéristiques acoustiques des résonances des voyelles sont affectées par la configuration d'autres articulateurs et pas seulement par le lieu d'articulation de la langue. En effet, une étude récente de Lawson et al. (2017) n'a pas trouvé de corrélation entre l'antériorité acoustique et l'antériorité articulatoire de /u/ dans plusieurs variétés d'anglais britannique. Les auteurs ont conclu que leur étude illustrait l'insuffisance d'une approche purement acoustique dans la recherche sur la variation vocalique diatopique.

Harrington et al. (2011) ont remarqué que les caractéristiques acoustiques d'un /u/ antériorisé pouvaient être la conséquence d'une position antérieure de la langue ou d'un manque d'arrondissement des lèvres. A partir des données d'articulographie électromagnétique, Harrington et al. (2011) ont démontré qu'en anglais standard du sud de l'Angleterre, ce qui distingue la voyelle /u/ de la voyelle /i/ n'est plus la position de la langue mais plutôt l'arrondissement des lèvres. Par conséquent, nous nous intéressons à cette voyelle parce que nous anticipons qu'elle sera particulièrement affectée par l'écarteur. Si c'est bien le cas, une compensation possible pourrait impliquer une postériorisation de la langue pour allonger la cavité antérieure. Les données échographiques de la langue vont nous permettre de juger dans un premier temps si la voyelle /u/ a un lieu d'articulation antérieur dans la production non perturbée. Nous pourrons ensuite comparer la position de la langue pour /u/ dans la parole normale avec celle produite pendant la parole perturbée. Nous comparerons enfin les données articulatoires de /u/ avec les données acoustiques.

# 3 Méthodologie

## 3.1 Participants et stimuli

Nous avions initialement enregistré cinq Britanniques à Paris, mais en raison de problèmes de visualisation des données échographiques chez une locutrice, nous présenterons les données de quatre locuteurs, dont un homme et trois femmes de 26, 50, 26, et 26 ans respectivement. Trois participants vivaient à Paris et parlaient le français (niveau B2/C1). Cependant, tous les locuteurs avaient vécu en Angleterre en milieu monolingue jusqu'à l'âge de 20 ans au moins et parlaient quotidiennement l'anglais. Une locutrice produisait un anglais standard du sud de l'Angleterre tandis que les autres venaient du nord d'Angleterre (Yorkshire et Lancashire).

Nous avons choisi des paires minimales de type /hVd/ où V représente une des onze monophtongues de l'anglais standard du sud de l'Angleterre (/iː, ɪ, ɛ, æ, ɜː, uː, ʊ, ʌ, ɔː, ɑː, ɒ/). Trois répétitions de chaque mot ont été produites en isolation avec et sans l'écarteur de lèvres. Les stimuli ont été présentés dans un ordre aléatoire par sujet et par condition avec les mots distracteurs. Afin de tester la dynamique de la mâchoire, à la fin de chaque condition, les participants ont produit la phrase « *hit the nail on the head* » trois fois ainsi que cinq répétitions de la syllabe /ba/ à un rythme régulier.

## 3.2 Matériels et procédure

Les participants portaient un écarteur de lèvres d'une largeur de 130 mm et d'une hauteur de 90 mm. Le signal acoustique a été capté par le biais d'un microphone à condensateur cardioïde (AKG Perception 120) relié à une carte-son externe (Presonus Audiobox), et a été numérisé directement au format PCM Windows, mono, avec une fréquence d'échantillonnage de 22 050 Hz et une quantification de 16 bits. Les données articulatoires ont été enregistrées avec un système échographique (Echo Blaster 120) avec une fréquence de ~60 images par seconde relié à une sonde de 5-8 MHz dans le plan sagittal médian. Le signal acoustique et articulatoire ont été enregistrés et synchronisés avec Articulate Assistant Advanced (AAA) (Articulate Instruments, 2014). Les participants portaient un casque ajustable de stabilisation de la sonde échographique afin d'éviter les mouvements de la tête par rapport à la sonde (Articulate Instruments, 2008).

Les données accélérométriques ont été enregistrées avec la plateforme BITalino (Guerreiro et al., 2013). Les données ont été acquises avec le logiciel Open Signals avec une fréquence d'échantillonnage de 1 000 Hz et une quantification de 10 bits. Le capteur était collé au menton des participants et mesurait simultanément l'accélération sur les axes X, Y et Z. Afin de synchroniser les données accélérométriques avec le signal acoustique, un bouton relié au BITalino permettait de déclencher et d'interrompre l'enregistrement dans AAA et d'inscrire l'événement sur une entrée numérique du BITalino pour permettre la postsynchronisation des données.

Nous avons effectué les enregistrements dans deux conditions : pendant la parole perturbée (P) et sans perturbation, donc pendant la parole normale (N). Deux participants ont commencé par la condition P et ont terminé par la condition N. On a inversé ces conditions pour les deux autres participants. Les participants étaient confortablement assis dans une chambre sourde. On leur a présenté les stimuli sur un écran externe. Ils ont été informés qu'il faudrait dire chaque mot/phrase qui apparait sur l'écran de la manière la plus naturelle possible après avoir entendu un bip. Chaque passation a duré moins de 30 minutes au total.

FIGURE 2 : Dispositif expérimental avec l'écarteur de lèvres, l'accéléromètre, et un casque de stabilisation relié à la sonde échographique

## 3.3 Analyse des données

### *3.3.1 Accéléromètre*

Les données continues brutes de l'axe vertical de l'accéléromètre ont été converties en *g*, synchronisées avec le début de chaque stimulus audio, et segmentées. L'offset des segments résultants a été éliminé et le signal a été rectifié. Nous avons ensuite calculé l'amplitude moyenne de chaque essai comme un indicateur de l'accélération.

### *3.3.2 Acoustique*

Le signal acoustique a été segmenté et étiqueté avec Praat (Boersma & Weenink, 2017). Les valeurs formantiques ont été extraites au milieu de la voyelle avec l'algorithme Burg dans Praat. Nous avons ajusté empiriquement les paramètres de l'algorithme pour obtenir une estimation satisfaisante des valeurs formantiques superposées au spectrogramme[2]. Pour définir si la voyelle /u/ était antériorisée, nous avons comparé sa valeur de F2 avec celle des voyelles périphériques /i/ et /o/. Si le F2 était plus proche du celui de /i/, nous avons conclu que /u/ était effectivement antériorisé.

### *3.3.3 Echographie de la langue*

Nous avons, dans un premier temps, visualisé d'une manière globale les contours de la langue pendant la production de chaque voyelle. Une image échographique a été sélectionnée pour chaque voyelle pour chaque locuteur avec et sans perturbation. Il s'agit de l'image la plus nette qui montre le contour de la langue dans sa position maximale avant le début du geste coronal de la consonne suivante (/d/). Nous avons utilisé AAA afin de détecter d'une façon quasi-automatique le contour de la langue de chaque image. Les erreurs de détection ont été corrigées à la main. Les contours ont été convertis sous forme de coordonnées dans un espace à deux dimensions (en millimètres).

Les images échographiques n'incluent que la surface de la langue et non pas des structures immobiles du conduit vocal, ce qui rend la tâche de quantification difficile car on n'a pas de point de

[2] Les scripts sont disponibles à https://tinyurl.com/hwv6a96

référence. Si la position de la sonde reste constante (à l'aide d'un casque de stabilisation par exemple), il est possible de quantifier les différentes positions de la langue. Nous avons opté pour la méthode de Scobbie et al. (2012), qui définit l'espace vocalique en fonction des coordonnées de deux voyelles périphériques, notamment /i/ et /o/. Pour caractériser la position relative de /u/, on dessine une tangente commune qui relie ces deux voyelles. Cette ligne de référence est ensuite traitée comme l'orientation horizontale de chaque locuteur pour déterminer le degré d'antériorisation de /u/. Si le /u/ est plus proche du /i/ que du /o/, nous concluons que le locuteur présente une antériorisation linguale de /u/, comme le cas présenté à la FIGURE 3.

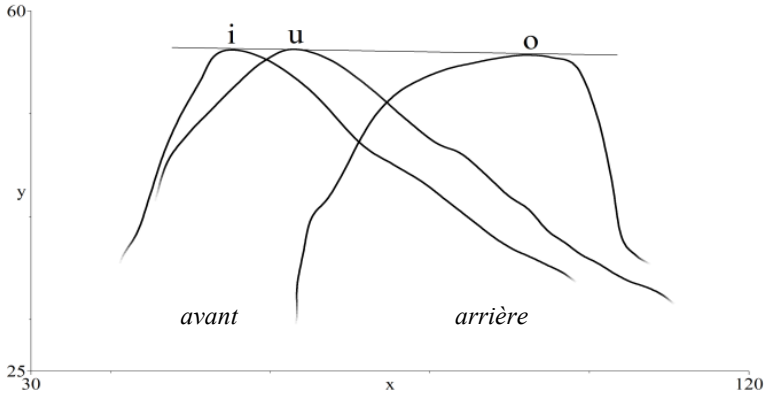



FIGURE 3 : L'utilisation d'une tangente commune pour définir la localisation des voyelles.

# 4 Résultats et interprétation

## 4.1 Accéléromètre : perturbation de la dynamique de la mâchoire

Nous avons calculé l'accélération moyenne comme mesure exploratoire de la perturbation de la dynamique de la mâchoire engendrée par l'écarteur. En moyenne, l'accélération était à 0,0741*g* dans la condition N et à 0,0433*g* dans la condition P avec un ratio P/N à 0,5792. Ce ratio indique que l'accélération baisse d'environ 40% lorsque la production est perturbée. Nous avons comparé les valeurs pour les deux stimuli : « ba ba ba ba ba » et « hit the nail on the head ». Les valeurs pour le premier étaient presque deux fois plus élevées que celles enregistrées pour le deuxième car la production de l'occlusive bilabiale nécessite un mouvement mandibulaire plus ample que celui occasionné par la plupart des consonnes du second. Comme plus haut, ces données sont influencées par la perturbation. Un développement logique de cette analyse très préliminaire consistera à examiner plus finement l'accélération induite par chaque son séparément.

## 4.2 Analyse acoustique

FIGURE 4 présente les modifications moyennes de l'espace vocalique après la perturbation pour chaque locuteur. L'espace vocalique de la locutrice 4 du sud est unique : c'est le seul qui présente une distinction entre les deux voyelles /ʊ/ et /ʌ/. On remarque également que sa production non-perturbée de /u/ est beaucoup plus antérieure que celle des autres locuteurs, avec une différence moyenne d'environ 962 Hz sur F2. Par conséquent, à partir des données acoustiques, on pourrait conclure que la locutrice du sud est la seule à présenter une antériorisation de /u/ dans la parole non-

perturbée. On y observe également une antériorisation de la voyelle /ʊ/, ce qu'on n'observe pas chez les participants du nord.

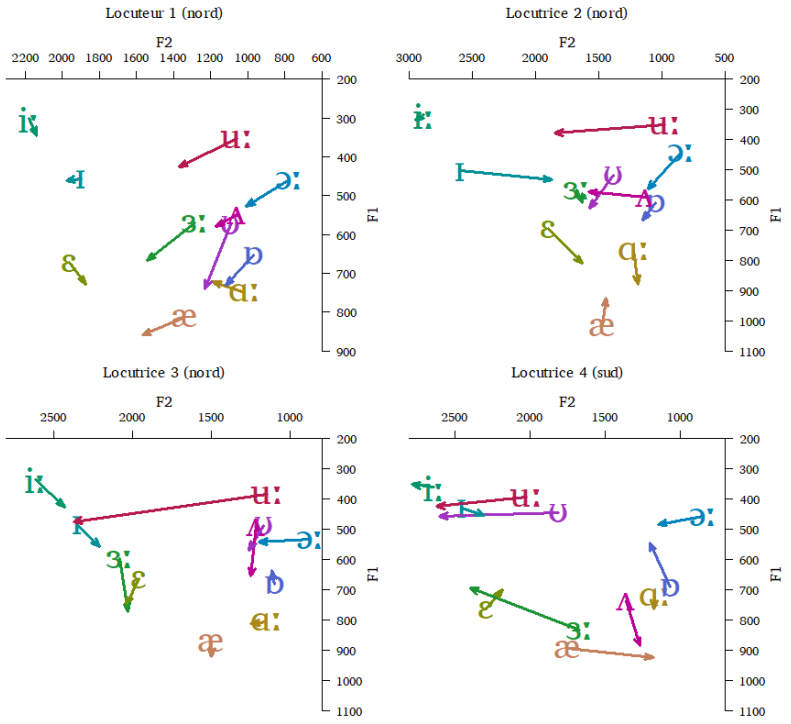



FIGURE 4 : Les modifications de l'espace vocalique après la perturbation

Nous avons observé beaucoup de variation parmi les locuteurs suite à la perturbation, qui découle probablement de l'effet de l'écarteur de lèvres sur des conduits vocaux de tailles différentes. Afin d'examiner quelles voyelles étaient particulièrement influencées par la perturbation, nous avons calculé la distance euclidienne dans F1-F2 en Bark entre les conditions N et P pour chaque voyelle. Chez les trois locuteurs du nord, la voyelle la plus affectée par la perturbation est /u/. L'effet est moins saillant pour cette voyelle chez la locutrice du sud. Ceci indique potentiellement une utilisation importante des lèvres pour la production habituelle de /u/ chez ces trois locuteurs, contrairement à la locutrice du sud, chez qui l'effet de perturbation est moins saillant pour /u/. Il semble donc que l'utilisation de la perturbation labiale puisse nous fournir les informations relatives à la variation sociophonétique : les locuteurs du nord étant plus affectés par la perturbation utilisent habituellement donc plus de labialisation que la locutrice du sud. On n'aurait pas pu arriver à cette conclusion en s'appuyant simplement sur les données acoustiques de la parole non-perturbée.

Dans les descriptions traditionnelles des voyelles d'anglais, il n'existe pas de voyelles qui se distinguent uniquement par l'arrondissement des lèvres, contrairement à d'autres systèmes vocaliques comme le français et l'allemand. Il est donc possible que l'arrondissement des lèvres ne soit pas très saillant en anglais. Dans une étude comparative des systèmes vocaliques de l'anglais, de l'espagnol, du français et de l'allemand, Delattre (1969) a remarqué que l'arrondissement employé dans les voyelles arrondies anglaises est moins important que celui produit des voyelles équivalentes dans les autres langues étudiées. Cependant, nos données indiquent que lorsqu'on empêche l'arrondissement avec une perturbation labiale, la voyelle /u/ peut occuper le même espace vocalique acoustique F1/F2 que /i/ (chez locutrices 3 et 4). Il semble donc que l'arrondissement soit plus pertinent pour la distinction /i-u/ que la littérature ne le suggère. Ceci est peut-être dû à l'antériorisation de /u/ : ce qui distingue /u/ de /i/ n'est pas nécessairement la position de la langue mais plutôt l'arrondissement des lèvres, comme Harrington et al. l'avaient indiqué (2011).

### 4.3 Echographie de la langue

La tangente commune qui relie les voyelles /i/ et /o/ nous a permis de juger si la voyelle /u/ était produite avec une antériorisation de la langue. Si la position maximale de la langue pendant la voyelle est plus proche de celle de /i/ que celle de /o/, on juge la voyelle /u/ comme étant antériorisée. Chez les locutrices 2-4, la voyelle /u/ est antériorisée dans leur parole normale. Chez les locuteurs 1, 3 et 4, la position maximale de la langue devient plus proche du /o/ pendant la parole perturbée, donc plus *postérieure* pour /u/. Nous concluons qu'il s'agit d'une stratégie de compensation articulatoire suite à la perturbation labiale. Cependant, nos données acoustiques montrent qu'avec le degré de postérisation observé, les locuteurs n'arrivent pas à obtenir la même cible acoustique que celle d'un /u/ non-perturbé. Une modélisation acoustique pourrait nous aider à juger si une compensation totale est en effet possible avec ce genre de perturbation labiale.

### 4.4 Corrélats acoustiques-articulatoires

Selon nos données acoustiques, seule la locutrice du sud (4) présente une antériorisation de /u/ dans sa parole normale. Cependant, les données articulatoires montrent que la langue est plus proche de la voyelle antérieure /i/ que de la voyelle postérieure /o/ chez les deux autres locutrices (2, 3), alors que leurs valeurs de F2 indiquent le contraire. Il semble que cette absence de corrélation entre les données acoustiques et articulatoires soit due à l'influence des lèvres sur les valeurs formantiques. Les locutrices 2-4 utilisent toutes les trois un lieu d'articulation antérieur pour /u/ mais seule la locutrice 4 présente des valeurs de F2 élevées, ce qui indique que les locutrices 2 et 4 utilisent potentiellement plus de labialisation (ce que les valeurs centrées réduites de F1 et F2 ont également indiqué). Les données acoustiques de la parole non-perturbée ne nous fourniraient pas ces informations, ce qui corrobore donc la conclusion de Lawson et al. (2017) selon laquelle une approche purement acoustique n'est pas nécessairement suffisante.

Les valeurs de F2 nous montrent que la production de /u/ devient plus antérieure dans la parole perturbée chez tous les locuteurs. Mais en regardant les données échographiques, trois locuteurs sur quatre (1, 2, 4) ont utilisé un lieu d'articulation plus *postérieur* dans cette condition. Cette postériorisation induite par la perturbation met donc en évidence une compensation articulatoire. Cette compensation linguale n'est pas observable à partir des données acoustiques, ce qui renforce à nouveau l'importance des données articulatoires dans les études sur la compensation.

## 5 Conclusion

Malgré la variation occasionnée par la perturbation, l'utilisation de l'écarteur de lèvres nous semble appropriée pour les études en phonétique, surtout pour éliminer entièrement l'influence des lèvres et pour restreindre les mouvements de la mâchoire. Notre étude a montré les signes d'une compensation articulatoire suite à la perturbation avec cet appareil. Nous sommes conscients que cette méthode présente certains inconvénients. L'appareil n'est pas modulable, ce qui serait pourtant nécessaire pour s'adapter à des conduits vocaux de tailles différentes. L'inconfort est évidemment un inconvénient et nous conseillons donc de l'utiliser sur des périodes brèves. Les futures études pourraient porter sur son effet sur d'autres systèmes vocaliques ou sur les consonnes labiales ([p], [b], [w], [ɪ], etc.). Une modélisation acoustique sera également bénéfique. Le résultat majeur et inattendu de cette étude porte sur la décorrélation entre les données acoustiques et articulatoires. Notre soulignons donc l'importance de données articulatoires dans les études sur l'effet de la perturbation articulatoire et plus généralement sur la production des voyelles.

# Références


ARTICULATE INSTRUMENTS (2008). Ultrasound stabilisation headset users manual, revision 1.4. Edinburgh: Articulate Instruments Ltd.

ARTICULATE INSTRUMENTS (2014). Articulate Assistant Advanced ultrasound module user manual, revision 2.16. Edinburgh: Articulate Instruments Ltd.

BOERSMA P., WEENINK D. (2017). Praat: doing phonetics by computer. Version 6.0.26.

DELATTRE P. (1969). An acoustic and articulatory study of vowel reduction in four languages. *International Review of Applied Linguistics in Language Teaching,* 7(4), 295-326.

FANT G. (1960). *Acoustic theory of speech production.* The Hague: Mouton.

GHIO A. (2007). Modélisation du conduit vocal. Dans P. AUZOU, V. ROLLAND, S. PINTO et C. OZSANCAK (dir.), *Les Dysarthries* (140-156). Marseille : Solal.

GUERREIRO J., MARTINS R., SILVA H., FRED A.L. (2013). BITalino-A Multimodal Platform for Physiological Computing. *ICINCO* 1, 500-506.

HARRINGTON J., KLEBER F., REUBOLD U. (2011). The contributions of the lips and the tongue to the diachronic fronting of high back vowels in Standard Southern British English. *The Journal of the International Phonetic Association* 41, 137-156.

LAWSON E., STUART-SMITH J., MILLS L. (2017). Using ultrasound to investigate articulatory variation in the GOOSE vowel in the British Isles. Actes de *Ultrafest VIII*, 27-28.

LI T., HONDA K., WEI J., DANG J. (2015). A lip protrusion mechanism examined by magnetic resonance imaging and finite element modeling. Actes de *18th ICPhS*, Glasgow.

PERRIER P. (2005). Control and representations in speech production. *ZAS Papers in Linguistics* 40, 109-132.

RIORDAN C.J. (1977). Control of vocal-tract length in speech. *The Journal of the Acoustical Society of America* 62(4), 998-1002.

SAVARIAUX C., PERRIER P., ORLIAGUET J.P. (1995). Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production. *The Journal of the Acoustical Society of America* 98(5), 2428-2442.

SAVARIAUX C., PERRIER P., ORLIAGUET J.P., SCHWARTZ J.L. (1999). Compensation strategies for the perturbation of French [u] using a lip tube. II. Perceptual analysis. *The Journal of the Acoustical Society of America* 106, 381-393.

SCOBBIE J.M., LAWSON E., STUART-SMITH J. (2012). Back to front: a socially-stratified ultrasound tongue imaging study of Scottish English /u/. *Rivista di Linguistica* 24(1), 103-148.

STEVENS K.N. (1989). On the quantal nature of speech. *Journal of phonetics* 17(1), 3-45.

# Analyse acoustique des occlusives produites par des jeunes locuteurs en dialecte wu de Suzhou

WANG Ning[1]
(1) Laboratoire LACITO, CNRS, 94801 Villejuif, France
wangning0531@gmail.com

## RESUME

Le système phonologique consonantique du wu ne comporte aucune opposition de voisement pour les occlusives. Phonétiquement, on constate que les occlusives sourdes /p, t, k/ en position initiale peuvent être perçues voisées et légèrement soufflées quand elles sont suivies d'une voyelle à ton bas. En position intervocalique d'une unité complexe dissyllabique, leur réalisation peut être voisée. Cet article examine la condition du soufflement et du voisement dans le dialecte wu de Suzhou où on observe qu'à l'initiale et en registre bas, les jeunes locuteurs maintiennent le soufflement qui se propage à partir du milieu de la réalisation de la voyelle jusqu'à la fin ; à l'intervocalique, le degré de voisement est effectivement plus élevé mais seulement dans un contexte lexical et tonal spécifique. Ce degré ne réalise pas le voisement à 100%. De plus, on note que la réalisation intervocalique peut varier entre [b, d, g] et [ɓ, ɗ, ɠ].

## ABSTRACT

**Acoustical analysis of stops produced by young speakers of Suzhou dialect**

The consonantal phonological system of wu has no opposition of voicing for stops. Phonetically, we find that the stops /p, t, k/ in initial position can be perceived voiced and slightly breathy when followed by a vowel with low tone. In intervocalic position of a complex disyllabic unit, the realization of /p, t, k/ can be voiced. This article examines the condition of the breathiness and the voicing in the dialect spoken in Suzhou and we observe that in initial and low register, the young speakers maintain the breathiness that start from the middle of the realization of the vowel until the end; in intervocalic position, the degree of voicing is actually higher but only in a specific lexical and tonal context. This degree doesn't arrive at 100%. Moreover, we note that the intervocalic realization of three stops can vary between [b, d, g] and [ɓ, ɗ, ɠ].

MOTS-CLES: langue wu, dialecte de Suzhou, tonal haut et bas, voix soufflée acoustique et physiologique, voisement.

KEYWORDS: wu language, Suzhou dialect, low vs. high tone, acoustic/physiological breathiness, voicing.

# 1 Introduction

Si l'opposition de voisement au niveau des occlusives est pertinente en français entre /p/-/b/, /t/-/d/ et /k/-/g/, le mandarin quant à lui ne connaît qu'une opposition d'aspiration et oppose : /p/-/p$^{h}$/, /t/-


10.21437/JEP.2018-53

/$t^h$/ et /k/-/$k^h$/. La langue wu parlée à l'est de la Chine appartient à la même famille et au même groupe que le mandarin. Il n'est donc pas étonnant de retrouver la même corrélation d'aspiration au niveau des phonèmes occlusifs qu'en mandarin. Toutefois, ces phonèmes rencontrent dans certains contextes des particularités phonétiques et acoustiques qui n'existent absolument pas en mandarin. Le wu connaît des variétés régionales dont les plus importantes sont celle de Shanghai et celle de Suzhou. Cette étude s'intéresse au dialecte de Suzhou qui possède sept tons : *yin ping* « 44 » ; *yang ping* « 223 » ; *yin shang* « 51 » ; *yin qu* « 523 » ; *yang qu* « 231 » ; *yin ru* « 43 » ; *yang ru* « 23 » (Xin, 2011). En fonction de leur registre de départ initial : le registre haut *yin* regroupe les tons dont la hauteur initiale est égale ou supérieure à 3 sur l'échelle de hauteur tonale, le registre bas *yang* regroupe ceux dont la valeur tonale initiale est inférieure à 3. D'où quatre tons hauts et trois tons bas (Chao, 1930). Dans la littérature, la consonne initiale suivie d'une voyelle dont le ton est à un registre bas est considérée comme voisée. Or la perception voisée n'est pas manifeste puisque le spectrogramme prouve que, physiologiquement, elle est réalisée sans voisement. La perception voisée serait due à une voix légèrement soufflée ou aspirée (Chao, 1930 ; Cao et *al*., 1992 ; Shi, 2009 ; Gao et *al*., 2012) mais cette hypothèse reste controversée. A l'intervocalique d'une unité composée, nous constatons une neutralisation tonale conformément à la règle de sandhi qui veut que dans ce contexte, les syllabes, autres que la première syllabe d'une unité polysyllabique, héritent du ton de la première syllabe. Ce ton se propage vers la droite et les syllabes qui subissent le sandhi perdent leurs tons originaux. Par conséquent, une consonne intervocalique suivie d'une voyelle dont le ton, normalement, est au registre bas subit un changement tonal ce qui cette fois lui permet de se réaliser physiologiquement comme voisée (Shi, 2009). Par exemple : 番 [$fe^{44}$] +茄 [$ka^{223}$] = 番茄 [$fe^{44}ga^{31}$] « échouer ».

Notre étude étudie ce phénomène en testant la production de jeunes locuteurs de wu vivant à Suzhou sur leur réalisation des phonèmes /p, t, k/ à l'initiale des morphèmes et à l'intervocalique d'unités complexes. L'objectif est d'en mesurer le *Voice Onset Time* (VOT) initial et le v-ratio intervocalique et d'appliquer la méthode acoustique (H1-H2 et HNR) pour confirmer ou infirmer nos hypothèses.

# 2 Analyse de la production des phonèmes occlusifs

## 2.1 Méthode

### 2.1.1 Participants

Quatre hommes et trois femmes de 20 à 29 ans (soit une moyenne de 23 ans) ont participé activement à cette étude. Tous sont locuteurs natifs de wu et ont une connaissance et une pratique élevée de leur langue maternelle.

### 2.1.2 Stimuli

Ils ont été amenés à oraliser vingt et une unités complexes (TABLE 1). Elles sont toutes à l'écrit des mots dissyllabiques de structure : $C_1V_1C_2V_2$ où C peut être /p/, /t/ ou /k/ et V /a/ ou /æ/. 1 indique la position initiale et 2 indique la position intervocalique. La présence de /æ/ s'explique par le faible nombre de morphèmes ayant en position intervocalique après /t/ la voyelle /a/. Le phonème /a/ est proposé car son F1 est relativement haut et de ce fait, il influence moins le premier harmonique H1 (le fondamental) et le deuxième harmonique H2 qui sont les deux indices importants à mesurer.

| **Initiale** | **Intervocalique** |
|---|---|
| 牌子 [pa$^{22}$tsy$^{33}$]« marque » | 失败 [seɦ$^{43}$ba$^{31}$] « échouer » |
| 败家 [pa$^{22}$ka$^{33}$] « faraud » | 雪白 [se$^{43}$baɦ$^{43}$] « blanc » |
| 白色 [paɦ$^{23}$sə$^{43}$] « blancheur » | 单摆 [te$^{44}$pa$^{31}$] « pendule » |
| 摆设 [pa$^{51}$seɦ$^{3}$] « décor » | 一百 [iəɦ$^{43}$paɦ$^{43}$]« cent » |
| 百万 [paɦ$^{43}$mɛ$^{31}$/vɛ$^{31}$] « million » | 大伯 [dəu$^{22}$paɦ$^{3}$] « oncle » |
| 爸爸 [pa$^{44}$pa$^{31}$] « père » | 葡萄 [beɦ$^{23}$dæ$^{31}$] « raisin » |
| 大学 [ta$^{22}$gɦoɦ$^{3}$] « université » | 海带 [hɛ$^{51}$ta$^{23}$] « laminaire » |
| 带鱼 [ta$^{44}$ng$^{31}$] « tricheur » | 绊倒 [poe$^{44}$tæ$^{31}$] « buter » |
| 茄子 [ka$^{22}$tsy$^{33}$] « aubergine » | 番茄 [fe$^{44}$ga$^{31}$] « tomate » |
| 加油 [ka$^{44}$yeu$^{33}$] « encourager » | 添加 [tɦie$^{44}$ka$^{31}$] « ajouter » |
| 格子 [kaɦ$^{43}$tsɿ$^{51}$] « carreaux » | |

TABLE 1 - Les 21 unités composées produites par les 7 locuteurs wu de Suzhou.

### 2.1.3 Enregistrement

L'enregistrement a été fait avec une carte son (*Cakewalk USB AudioCapture*) et le *MicroMic AKG C5201* dans un studio professionnel à Suzhou en 2017. Les fichiers audios ont été enregistrés via le logiciel *Adobe Audition CC 2017*. Les participants ont fait les enregistrements assis pour leur confort mais surtout pour éliminer les balancements éventuels. Les locuteurs devaient prononcer avec un débit normal chaque mot deux fois, soit un total de 294 (21 termes x 2 fois x 7 locuteurs) stimuli. Tous les locuteurs n'ont pas produit les 21 mots, d'où seulement 228 stimuli d'enregistrés. Une session d'entraînement s'est déroulée avant l'enregistrement.

### 2.1.4 Paramètres acoustiques mesurés

En premier lieu, nous avons mesuré le *Voice Onset Time* (VOT) (Cho, 1999 et Lisker et *al*., 1964) qui correspond à la durée entre le relâchement consonantique et le début des pulsations périodiques glottiques. Il existe trois valeurs possibles : 1) le VOT est égal à zéro pour les occlusives non-voisées et non-aspirées ; 2) le VOT est négatif pour les occlusives voisées ; 3) le VOT est positif pour les occlusives aspirées.

À l'intervocalique, nous avons ensuite mesuré le v-ratio qui est le pourcentage de voisement pendant la durée de la consonne (Hallé et *al*., 2011) et cela pour deux raisons : 1) Sachant que la mesure du VOT ne se fait qu'en position initiale, le v-ratio propose une quantification du degré de voisement dès lors que la durée de la consonne peut être déterminée, ce qui est le cas à intervocalique ; 2) Le *Voice Termination Time* (V.T.T) (Agnello, 1975) marquant le résidu de la voyelle précédente pendant l'occlusion intervocalique (FIGURE 3, intervalle en vert clair) peut être pris en compte pour compléter le calcul à partir du passage transitoire. Cette durée est également appelée *edge vibrations* par Lisker & Abramson (1964) ou *bleed* par Davidson (2016).

La mesure de H1-H2 se base sur le paramètre acoustique *spectral tilt* (Fant, 1982 ; Bickley, 1982 ; Gordon et *al*., 2001) : au sein d'un segment voisé, pour une voix soufflée, le signal de la source voisée, correspondant à une onde quasi sinusoïdale, renforce l'amplitude de H1. Par conséquent, la différence d'amplitude entre le premier et le second harmonique est plus large pour un segment aspiré.

L'*Harmonics to Noise Ratio* (HNR) estime la proportion de l'intensité du signal harmonique en fonction du bruit (Krom, 1993, Murphy, 1999). Une voix soufflée étant accompagnée de davantage de bruit turbulent, le HNR est nécessairement plus faible que pour une voix non-soufflée.

Pour faire ces analyses, nous avons été amenés à modifier le script de Landron (2015) pour le logiciel *Praat* (Boersma, 2011) servant à mesurer le VOT et le v-ratio. Le programme *VoiceSauce* de Shue (2017) a été utilisé pour mesurer le HNR et le H1-H2. La segmentation et l'alignement ont été faits et vérifiés manuellement.

La segmentation a été faite par le biais des conventions suivantes (voir FIGURE 1) :

a) Pour $C_1$, le début est arbitraire puisqu'aucune information sur le début de l'occlusion n'est donnée sur le spectrogramme et l'oscillogramme. Le seul but de fixer une frontière initiale permet de faire fonctionner le script.

b) Pour $C_2$, le début a été déterminée après la fin de $F_2$ de la voyelle précédente, l'endroit où la dernière courbe harmonique traverse la ligne zéro en descendant. Pour la fin de la réalisation de $C_1$ et $C_2$, nous avons choisi le moment où le signal traverse, en montant, la ligne zéro.

c) Pour la mesure du v-ratio, notre script calcule des séquences (*frame*). Il calcule le pourcentage de séquences voisées par rapport à toutes les séquences délimitées (plus de 27 ms). Le v-ratio est calculé sur toute la durée de la consonne. Par contre, la mesure de v-ratio n'indique pas où se situe le voisement. Pour tous types de $C_2$, nous nous attendons à un v-ratio positif en tenant compte de la *Voice Termination Time* (V.T.T).

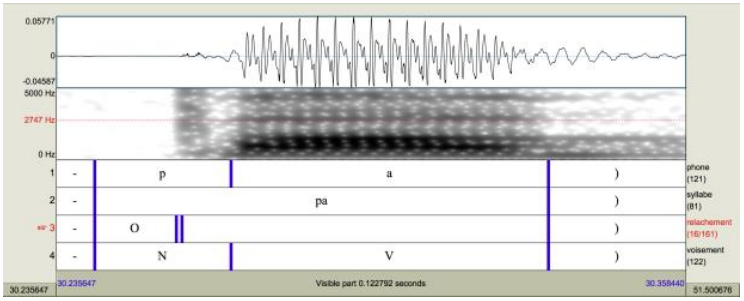



FIGURE 1 : Les tires 1,3 et 4 sont utilisées pour le calcul du VOT et du v-ratio. Sur la troisième tire, O signifie « Occlusion », R qui n'est pas visible ici signifie « Relâchement » et se situe dans l'espace très étroit après O. Sur la quatrième tire, N signifie « phase non-voisée » et V « phase voisée ». Le tiret - a été ajouté avant $C_1$ pour symboliser le silence tandis que la parenthèse fermante) remplace $C_2$. La délimitation arbitraire du début de l'occlusion pour $C_1$ est requise pour le calcul du VOT.

## 2.2 Résultats

### 2.2.1 HNR et H1-H2

Pour simplifier la présentation des syllabes avec différents tons, quatre étiquettes sont proposées : $C_1V_1^B$ pour un ton bas ; $C_1V_1^H$ pour un ton haut ; $C_2V_2^{(B)}$ pour un ton inhérent bas mais modifié par sandhi ; $C_2V_2^{(H)}$ pour un ton inhérent haut modifié par sandhi.

Les deux indices ont été mesurés au début (1/3), au milieu (2/3) et à la fin (3/3) de la réalisation des voyelles initiale et intervocalique. Pour H1-H2, la moyenne obtenue pour les trois intervalles a montré que $V_1^B$ était manifestement plus soufflée que lors de la production de $V_1^H$ (TABLE 2). Par contre, la caractéristique soufflée ne joue pas de rôle distinctif à l'intervocalique.

À propos du HNR, la mesure a été faite entre 0 à 3500 Hz. Le résultat montre que seul le HNR05 (entre 0-500 Hz) à l'initiale est significatif au milieu et à la fin de la réalisation (TABLE 2). C'est-à-dire que le HNR de $V_1^B$ est moins élevé que celui de $V_1^H$ et la différence se trouve au-dessus du $F_1$ pour $V_1$ qui est soit /a/ soit /æ/ ($F_1$ se situe entre 168Hz et 242Hz). Paradoxalement, il n'y pas de différence au début alors qu'elle devrait subir une influence transitoire plus forte.

| **Mesure (dB)** (n=58,54) | **1/3** | **2/3** | **3/3** |
|---|---|---|---|
| $H_1$-$H_2$ | 1.45* | 1.78* | 2.32** |
| HNR05 | -0.88 | -3.97* | -6.08** |

TABLE 2 : La différence de $V_1^B$-$V_1^H$ en fonction des mesures H1-H2 et HNR05 en trois points de la voyelle. Signification : * pour p <.05, ** pour p <.01.

### 2.2.2 VOT et v-ratio

Le VOT de $C_1V_1^B$ est plus court que celui de $C_1V_1^H$ mais la différence n'est pas significative. Certes, nous avons remarqué quelques réalisations pré-voisées chez un locuteur, mais la majorité des locuteurs ne l'a pas produit à l'initiale. La moyenne du VOT reste positive pour $C_1V_1^B$ et $C_1V_1^H$ (TABLE 3). Ainsi, l'utilisation du terme « sonore » ne correspond pas ici à la notion traditionnelle de voisement.

Le v-ratio de $V_2^{(B)}$ s'avère plus élevé que celui de $V_2^{(H)}$. (TABLE 3). Cela suggère une réelle opposition au niveau du degré de voisement. Mais ce qui est remarquable c'est le taux de v-ratio assez élevé qui se manifeste lors de la production de $C_2V_2^{(H)}$. Il est probable que les $C_2$ étaient largement pré-voisées en structure $C_2V_2^{(H)}$ même si ce type de $C_2$ a été considéré comme sourd dans les études précédentes (Chao 1935, Cao 1987, Cao et *al*., 1992, Hu, 2001, Shi, 2009). De plus, cette proportion considérable pourrait s'interpréter par la prise en compte du VTT (cf. 2.1.4) dans un contexte intervocalique facilitant la poursuite des vibrations des plis vocaux.

Par ailleurs, en contexte $C_2V_2^{(B)}$, certaines réalisations phonétiques des consonnes sont implosives ainsi /p/ est réalisé [ɓ] avec des amplitudes de voisement très importantes (FIGURE 2) ce qui n'est absolument pas attesté pour la réalisation de /p/ en contexte $C_2V_2^{(H)}$ (FIGURE 3).

| VOT (ms) | | v-ratio % | |
|---|---|---|---|
| $C_1V_1^B$(n=58) | $C_1V_1^H$(n=54) | $C_2V_2^{(B)}$(n=56) | $C_2V_2^{(H)}$(n=60) |
| 10.97 | 15.94 | 88.25** | 73.92** |

TABLE 3 : À gauche, la moyenne du VOT de $C_1V_1^B$ et $C_1V_1^H$. À droite, la moyenne du v-ratio entre $C_2V_2^{(B)}$ et $C_2V_2^{(H)}$. Significations : * pour p <.05, ** pour p <.01.

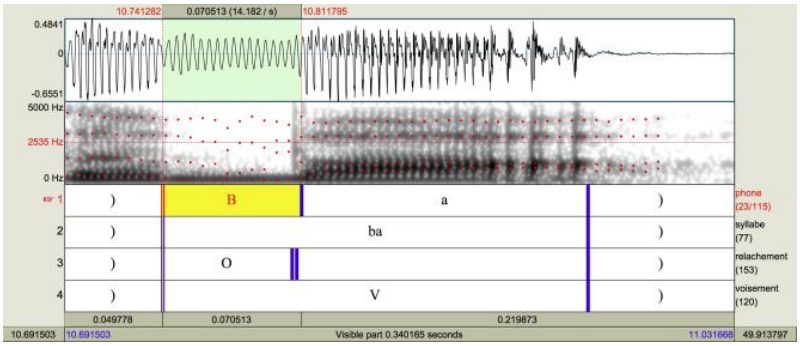



FIGURE 2 : Réalisation de [ɓ] par une locutrice wu de Suzhou. B noté dans la tire « phone » sert à indiquer une consonne intervocalique

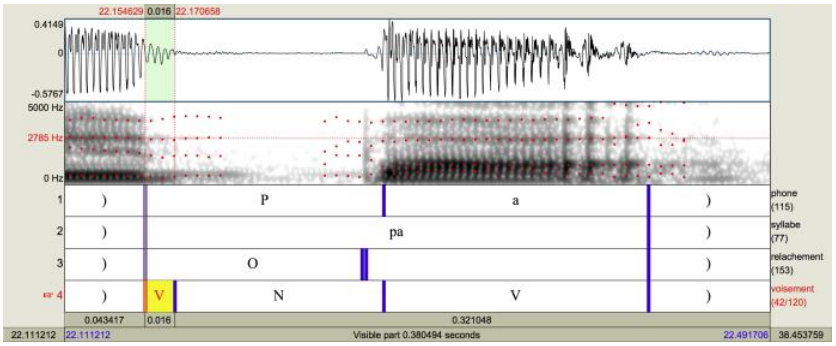



FIGURE 3: Le phonème /p/ est réalisé [p] par une locutrice wu de Suzhou. P note dans la tire « phone » la consonne intervocalique.

# 3 Discussion et conclusion

Dans ce travail, nous avons exploré les caractéristiques acoustiques des phonèmes occlusifs du dialecte wu de Suzhou. Nos résultats montrent que :

1) À l'initiale d'un morphème, le VOT ne sert pas à distinguer une série sonore dû à ses valeurs positives comme l'avaient déjà confirmé les scientifiques dans la littérature (Cao et *al*., 1992 ; Shi, 2009 ; Gao et *al*., 2012). En revanche, la différence des deux indices H1-H2 et HNR pourrait suggérer une plus grande proportion d'air turbulent lorsque $V_1$ porte un ton bas. D'où le terme de léger soufflement ou d'aspiration légère largement répandu.

Mais les mesures de ces deux facteurs n'ont pas donné des résultats homogènes notamment lorsque l'on cherche à préciser à partir de quel moment le soufflement commence à intervenir. Le résultat de HNR, qui a montré une différence significative entre $V_1^B$ et $V_1^H$ à partir des 2/3 de la durée vocalique, est à l'antipode de celui de H1-H2, qui a pu relever une différence significative dès le début de la voyelle. En fait, ces deux mesures s'intéressent à différents composants acoustiques : 1) La mesure de H1-H2 propose une voix plus soufflée quand le signal de la source voisée correspondant à une onde quasi sinusoïdale renforce l'amplitude de H1, d'où H1-H2 plus saillant. 2) La mesure de HNR dans notre étude, estime la proportion du bruit turbulent de la partie vocalique, d'où le HNR plus bas pour une voyelle plus soufflée. Selon une étude précédente (voir Gao et *al*., 2012) sur le dialecte de Shanghai produit par des jeunes locuteurs (moyenne 26 ans), le H1-H2 permet de localiser le soufflement à l'*onset* de la voyelle et l'effet se propage jusqu'au milieu. Par contre, elle n'a pas fourni les données du HNR. Nos résultats hétérogènes sur le début de l'apparition du soufflement nous invitent à réfléchir au niveau physiologique : a) Sachant qu'au niveau physiologique, une grande proportion de turbulence d'air sortant pendant l'ouverture des plis vocaux demande une abduction écartée des plis vocaux suite à l'action crico-aryténoïdienne. « Aspiration is presumably a result of a somewhat open glottis » (Abramson et *al*., 2017 : 76), est-ce qu'au début de la voyelle, la manœuvre laryngale compliquée est capable d'effectuer une abduction assez grande afin de créer une voyelle plus soufflée ? b) Sachant que ce soufflement se produit dans le contexte où le registre tonal est relativement bas, serait-il possible que la caractéristique soufflée

se stabilise et se renforce après le passage transitoire entre consonne et voyelle ? Dès que les plis vocaux courts et épais sont configurés en position stable et relâchée par le muscle crico-thyroïdien donnant ainsi un accès facile à la partie basse et médium de la voix, le soufflement se manifesterait.

2) À l'intervocalique d'une unité complexe dissyllabique, nous n'avons pas remarqué de différence significative pour H1-H2 et HNR entre $C_2V_2^{(B)}$ et $C_2V_2^{(H)}$. Par contre, une réalisation phonétique voisée a été constatée pour $C_2$ dans les deux contextes et le degré de voisement a une corrélation forte avec le ton intrinsèquement bas et modifié par sandhi de la $V_2$. Par contre, nous n'avons pas obtenu la moyenne concernant le v-ratio pour $C_2V_2^{(B)}$ la hauteur attendue de 100% montrant que $C_2$ est toujours voisée à l'intervocalique (Cao et *al.*, 1992 ; Shi, 2009 ; Gao et *al.*, 2012). La FIGURE 4 montre un /p/ partiellement voisé repéré dans plusieurs réalisations de trois locuteurs. De plus, la partie encadrée ne devrait pas être une simple pause. Il est fort probable que cette partie puisse se traduire par une série d'ouverture et de fermeture glottique irrégulière entre la diminution considérable des amplitudes suivie de la voyelle précédente ([e] dans cette séquence) et le pré-voisement de la consonne intervocalique.

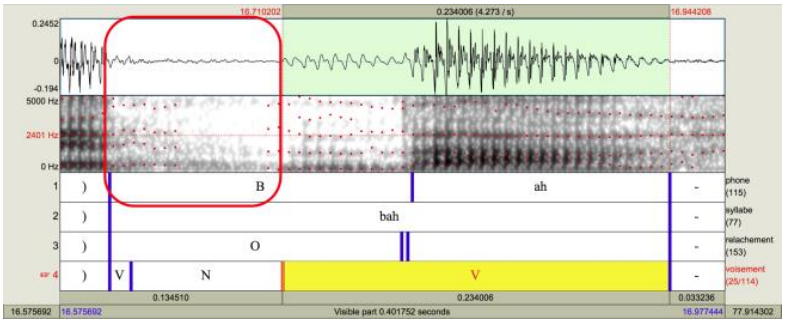



FIGURE 4: Réalisation partielle d'un [b] par une locutrice wu de Suzhou. B noté dans la tire « phone » indique la consonne intervocalique. La partie encadrée pourrait correspondre à une série d'ouverture et de fermeture glottique avant que le voisement ne se stabilise.

En résumé, nous pouvons donc déduire qu'à l'initiale, la caractéristique soufflée se manifeste d'une manière robuste dans le dialecte de Suzhou malgré le contact intense avec le mandarin. Cette particularité phonétique a été confirmée par les données acoustiques testées à Shanghai (Gao et *al.*, 2012). Donc, le wu resterait en synchronie et d'un point de vue phonétique une langue sinitique différente du mandarin. De plus, à l'intervocalique, on a observé une augmentation du degré de voisement chez les jeunes locuteurs ce qui les conduit à produire des implosives [ɓ, ɗ, ɠ] en contexte $C_2V_2^{(B)}$. Par contre, il nous reste encore des questions à résoudre notamment à propos du mécanisme complexe permettant le pré-voisement partiel et le mécanisme amenant le soufflement initial. Pour y répondre, nous recourrons à EVA2$^{TM}$ (*Computerised Vocal Assessment*, SQLab) qui quantifie le débit d'air et la pression intra-orale dans nos prochaines études.

# Références


ABRAMSON A.S., Whalen D.H. (2017). Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. *Journal of Phonetics* 63, 75-86.

AGNELLO J.G. (1975). Voice onset and voice termination features of stutterers. *Vocal tract dynamics and dysfluency: the proceedings of the first annual Hayes Martin Conference on Vocal Tract Dynamics*. New York: Webster, L.M. et Furst, L.C..

BICKLEY C.A. (1982). Acoustic analysis and perception of breathy vowels. *MIT Speech communication group working papers* 1, 71-81.

BOERSMA P. (2011). Praat, a system for doing phonetics by computer. *Glot International* 5(9/10), 341-345.

CAO J., IAN M. (1992). An exploration of phonation types in Wu Dialects of Chinese. *Journal of Phonetics*. 20,79-92.

CHAO Y. (1930). A System of Tone Letters. *Le Maitre Phonétique* 45, 24-27.

CHAO Y. (1935). *Xiandai wuyu de yanjiu [L'étude du dialecte Wu contemporain]*. Beijing. Shangwu Yinshuguan [La maison d'édition commerciale].

CHO T. LADEFOGED P. (1999). Variation and universals in VOT: evidence from 18 languages. *Journal of Phonetics* 27, 207-229.

DAVIDSON, L. (2016). Variability in the implementation of voicing in American English obstruents., *Journal of Phonetics* 54, 35–50.

FANT G. (1982). Preliminaries to analysis of the human voice source. *Quarterly Progress and Status Report, Speech Transmission Laboratory, KTH, Stockholm, Sweden*. 23(4), 1-27.

GORDO M., LADEFOGED P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics*. 29(4), 383-406.

GAO J., HALLÉ P. (2012). Caractérisation acoustique des obstruantes phonologiquement voisées du dialecte de Shanghai. Actes de *les Journées d'Études sur la Parole*, 57-64.

HALLÉ P., ADDA-DECKER M. (2011). Voice assimilation in French obstruents: A gradient or a categorical process? *Tones and features: A festschrift for Nick Clements*. De Gruyter, 149-175.

KROM G.D. A Cepstrum-Based Technique for Determining a Harmonics-to-Noise Ratio in Speech Signals. (1993). *Journal of Speech, Language, and Hearing Research*. 36(2), 254-266.

LANDRON S. (2017). *L'opposition de voisement des occlusives orales du français par des locuteurs taïwanais*. Thèse en phonétique pour le grade de Docteur. Paris. Université Paris Nouvelle Sorbonne.

LISKER L., ABRAMSON A.S. (1964). A cross-language study of voicing in initial stops: acoustical measurements. 20 (3), 384-422.

MURPHY P.J. (1999) Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis. *The Journal of the Acoustical Society of America*. 105(5), 2866-2881.

SHI F. (2009). *Shiyan yuyinxue tanjiu [L'exploration de la phonologie expérimentale]*. Beijing. Beijing Daxue Chubanshe [Peking University Press].

SHUE Y. (2010). The voice source in speech production: Data, analysis and models. *UCLA dissertation*.

# Caractériser la distinctivité du système vocalique des locuteurs

Christine Meunier [1] & Alain Ghio[1]
(1) Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
christine.meunier@lpl-aix.fr, alain.ghio@lpl-aix.fr

## RESUME

L'objectif de notre étude est de caractériser les locuteurs du français grâce à un indice de distinctivité lors de la production de voyelles en parole spontanée. Cette distinctivité est le plus souvent établie selon la dispersion de l'espace vocalique. Des travaux précédents (Huet & Harmegnies, 2000) ont proposé un indice plus dynamique prenant en compte le rapport entre la dispersion de l'ensemble des voyelles du système et la dispersion moyenne de chaque voyelle dans sa catégorie. Nous nous inspirons de ces travaux pour proposer un indice de distinctivité (ID) en vue d'établir des profils de locuteurs. Nos premiers résultats confirment des différences interlocuteurs. L'indice lui-même n'est pas toujours en lien avec la dispersion globale du système et permet de mettre en évidence une interaction plus fine entre voyelle et système. Suite à cette première étape nous envisageons d'évaluer cet ID selon différents facteurs (langue, type de parole, populations pathologiques).

## ABSTRACT

**The characterization of the distinctivity in speakers' vowel production.**
The objective of our study is to characterize the French speakers thanks to a cue of distinctiveness in the production of vowel in spontaneous speech. Distinctiveness is most often derived from the dispersion of vowel space. Previous work (Huet & Harmegnies, 2000) has proposed a more dynamic cue taking into account the relationship between the dispersion of the whole vowels of the system and the average dispersion of each vowel in its category. To go on with this view we propose a cue of distinctiveness (ID) in order to provide speakers' profiles. Our first results confirm differences between speakers. The cue itself is not always related to the overall dispersion of the system but highlights a more precise interaction between the vowel and the system. Following this first step, we plan to evaluate this ID according to different factors (language, type of speech, pathological populations).

MOTS-CLES : distinctivité ; acoustique; voyelles; parole spontanée
KEYWORDS: distinctiveness ; acoustics ; vowels ; spontaneous speech

## 1 Introduction

L'objectif de notre étude est de caractériser les locuteurs du français au travers d'un indice de distinctivité lors de la production de voyelle en parole spontanée. L'organisation des réalisations vocaliques a depuis longtemps suscité l'intérêt des chercheurs au travers de plusieurs objectifs : la comparaison des langues, les variations dues à la situation de parole ou encore la comparaison de populations (saines/pathologiques, L1/L2). D'un point de vue méthodologique, le système vocalique présente l'avantage de faire apparaître des variations graduelles au sein d'un même mode de


10.21437/JEP.2018-54

production, ce qui n'est pas le cas pour les consonnes. A cet égard, la *centralisation* du système vocalique est apparue comme une conséquence des facteurs énumérés ci-dessus. Notamment (Smiljanić & Bradlow, 2009) précise que l'hyper-articulation des voyelles, associée à une parole « claire », augmente la distance F1-F2 et occasionne moins de chevauchement entre les réalisations de chaque catégorie de voyelle, ce qui les rend plus distinctes. Toutefois, l'hyper-articulation des voyelles peut être reliée à deux dimensions : une augmentation de l'espace vocalique global (du système) et/ou une réduction de l'espace de production de chaque catégorie de voyelle. Huet & Harmegnies (2000) propose un indice *Phi* permettant de de calculer le rapport entre la variation à l'intérieur du système ($CM_{inter}$) et la variation moyenne de chaque catégorie de voyelle ($CM_{intra}$). Cet indice a permis aux auteurs de montrer que la parole spontanée induisait un indice bien plus faible que des situations de lecture chez un même locuteur. C'est donc sur la base de cet indice que nous souhaiterions apporter une contribution concernant la comparaison des locuteurs en parole conversationnelle en français.

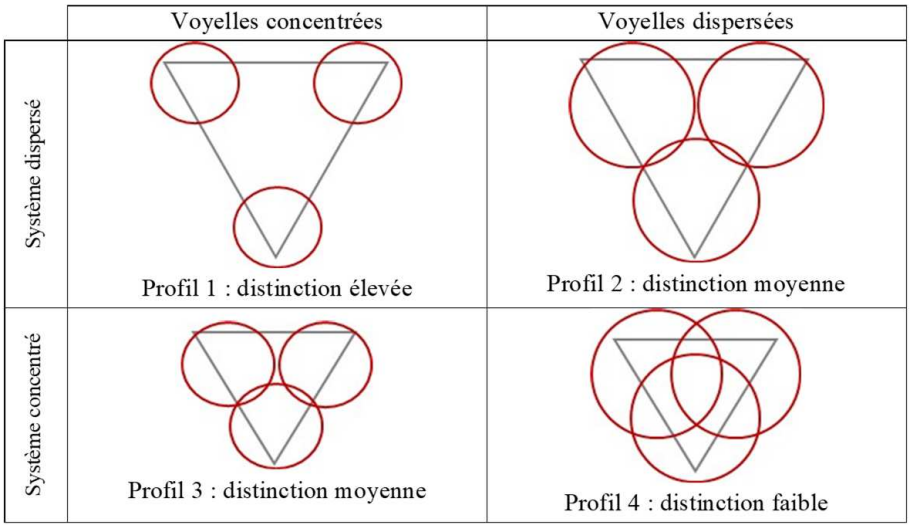



Figure 1 : Représentation schématique de la production des voyelles selon deux paramètres de dispersion: celui du système vocalique et celui de chaque catégorie de voyelle. Quatre profils peuvent être proposés suggérant des confusions plus ou moins importantes entre les voyelles.

L'interaction entre la dispersion/concentration du système vocalique et celles des catégories de voyelle peut être représentée schématiquement (Figure 1). Nous observons ici que le degré de distinctivité (donc la possibilité ou non de chevauchement des réalisations vocaliques) est fonction de ces deux facteurs. Ces profils représentent des hypothèses très caricaturales et nous envisageons évidemment que ces deux paramètres évoluent de façon corrélée c'est-à-dire qu'une concentration du système pourrait entraîner une concentration des voyelles. Toutefois, ça n'est pas ce qui a été observé dans un changement de situation de parole : la lecture de mots isolés serait plutôt similaire au profil 1 alors que le passage à une parole plus enchaînée engendrerait un profil 4 (Meunier, Espesser, & Frenck-Mestre, 2006). Ce que nous cherchons, dans cette première étape, est de faire apparaître des différences interlocuteurs dans un type de parole très relâché.

Pour ce faire, et à partir des travaux de (Huet & Harmegnies, 2000), nous avons calculé un Indice de Distinctivité (ID) permettant d'exprimer le rapport entre la dispersion du système et celle des voyelles chez des locuteurs de façon à obtenir des profils différents. A plus long terme, nous envisageons plusieurs types de comparaison : locuteurs de différentes langues (et donc systèmes vocaliques variés), locuteurs dans des situations de parole différentes (Huet & Harmegnies, 2000)

et locuteurs affectés par des pathologies de la parole (comme l'ont présenté Audibert & Fougeron, 2012).

# 2 Méthodologie

## 2.1 Corpus

Les analyses ont porté sur un corpus de parole conversationnelle, le *Corpus of Interactional Data* CID, (Bertrand et al., 2008). Ce corpus, enregistré en 2003, comprend des enregistrements audio et vidéo de dialogues spontanés entre des locuteurs français natifs (8 conversations d'une heure chacune entre deux locuteurs, soit 16 locuteurs, 10 femmes et 6 hommes). Dans chaque dialogue, les locuteurs entretenaient une conversation familière. Pour notre étude, nous avons sélectionné 10 locuteurs[1] : 5 femmes (AB, AC, BX, LL, ML) et 5 hommes (AG, AP, EB, LJ, SR). L'ensemble du corpus a bénéficié d'une transcription orthographique enrichie. Cette transcription a ensuite été phonétisée, puis alignée automatiquement de façon à obtenir une annotation phonétique (Bertrand et al., 2008).

Notre analyse porte exclusivement sur les voyelles orales. Les voyelles dont la durée était inférieure à 30ms ou supérieure à 300ms ont été exclues de nos analyses. En effet, les voyelles très courtes sont souvent (même si ça n'est pas systématiquement) le produit d'erreurs d'alignement dus à des omissions ou réductions non perçues pas les transcripteurs. De même, les voyelles extra-longues incluent très souvent des pauses remplies que nous ne souhaitions pas inclure dans cette étude. Enfin, un filtre établi par (Gendrot & Adda-Decker, 2005) a été appliqué de façon à exclure les valeurs de formant aberrantes dues à des erreurs de détection. En conséquence, un total de 37452 voyelles a été analysé dans notre étude (Table 1).

| Voyelles | AB | AC | AG | AP | BX | EB | LJ | LL | ML | SR | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| @ | 510 | 453 | 554 | 607 | 399 | 570 | 513 | 309 | 480 | 609 | 5004 |
| A | 925 | 959 | 1057 | 852 | 880 | 654 | 1087 | 473 | 1069 | 873 | 8829 |
| e | 1530 | 1278 | 1398 | 1303 | 1303 | 1106 | 1427 | 606 | 1289 | 1306 | 12546 |
| i | 716 | 449 | 437 | 525 | 482 | 450 | 697 | 318 | 528 | 484 | 5086 |
| o | 298 | 207 | 235 | 236 | 273 | 183 | 326 | 120 | 288 | 263 | 2429 |
| u | 132 | 118 | 105 | 119 | 116 | 41 | 115 | 60 | 129 | 87 | 1022 |
| y | 293 | 205 | 338 | 310 | 158 | 249 | 323 | 117 | 262 | 281 | 2536 |
| Total | 4404 | 3669 | 4124 | 3952 | 3611 | 3253 | 4488 | 2003 | 4045 | 3903 | **37452** |

Table 1: nombre de voyelles étudiées par locuteur et par voyelle

On constate que le nombre total de voyelles produites est très variable selon le locuteur (2003 pour LL et 4488 pour LJ) et est fonction du temps de prise de parole dans le dialogue pour chaque locuteur. De même, on observe que l'effectif de chaque catégorie de voyelle est très hétérogène. La voyelle /e/ est 12 fois plus représentée que la voyelle /u/. Cela tient à deux facteurs : le premier, bien évidemment, est la fréquence des voyelles dans le lexique, ainsi que la fréquence du lexique dans le

[1] Ces 10 locuteurs ont été sélectionnés pour deux raisons : 1/ pour une partie d'entre eux, nous disposons de corpus de lecture (mots et textes) avec lequel des comparaisons sont envisagées ; 2/ pour une autre partie, nous disposons d'annotations phonétiques fines (alignement corrigé par un expert humain) qui nous permettent des analyses plus poussées.

discours ; le deuxième réside dans le fait que le phonétiseur ne fait pas de distinction pour les voyelles moyennes qui sont regroupées en archiphonèmes /e/ (/e/, /ɛ/), /o/ (/o/, /ɔ/) et @ (/ə/, /ø/, /œ/). Quoiqu'il en soit, on observe que le système du français montre une majorité de voyelles antérieures (ou centrale). Cette tendance est accentuée par une sur-représentation de ces voyelles dans le discours. Elle n'est pas anodine pour le calcul du centre de gravité du système vocalique des locuteurs qui pourra ainsi tendre vers l'avant. Nous aurons à prendre en compte ce phénomène dans l'analyse des données (voir section ci-dessous)

## 2.2 Mesures

Les trois premiers formants ont été estimés automatiquement à l'aide d'une méthode de prédiction linéaire (autocorrélation) avec un algorithme Viterbi de façon à sélectionner les meilleurs candidats en imposant une contrainte de continuité fréquentielle (ESPS package, Entropic, 1997). Par la suite les mesures en Hertz des trois formants ont été converties en Bark selon la formule de Traunmüller (1990). En effet, dans la mesure où notre étude porte sur une analyse des distances euclidiennes entre chaque voyelle et son barycentre pour les trois formants, il nous a semblé important d'homogénéiser au mieux chacune des dimensions de l'espace. En effet, si on considère les variations de F1 de 348 à 685 Hz pour les femmes (Gendrot & Adda-Decker, 2005), cela représente une dynamique de 337Hz mais seulement 2.9 Barks d'écart. Si on considère les variations de F2 de 1140 à 2365 Hz, cela représente 1225 Hz d'écart (4 fois plus que pour F1) mais seulement 4.8 Barks. Enfin, pour F3, les auteurs mesurent des variations de 2687 à 3130 Hz ($\Delta$F=443 Hz) ce qui représente environ 1 Bark. L'homogénéité n'est pas parfaite mais bien meilleure en Bark qui, nous le rappelons, est une échelle psychoacoustique imitant les mécanismes de la perception humaine, notamment de distinctivité, ce qui est parfaitement cohérent avec notre travail.

Pour chaque locuteur et pour F1, F2 et F3, nous avons calculé: 1/ la moyenne en Bark des valeurs de chaque catégorie de voyelle (**MOY_VOY**) et 2/ la moyenne en Bark pour l'ensemble des voyelles produites (**MOY_SYS).** Ces moyennes nous permettent de calculer, pour chaque locuteur, les distances euclidiennes en trois dimensions (F1, F2 et F3) entre la valeur de chaque voyelle et la moyenne 1/ de sa catégorie et 2/ du système.

### *2.2.1 Calcul des Distances Euclidiennes par catégorie de voyelle (DE_3D_VOY)*[2]

DE_3D_VOY représente la dispersion des voyelles par rapport au centre de gravité de leur catégorie (dispersion des voyelles). Elle est obtenu en appliquant la formule suivante :

$$\sqrt{(F1v - F1V)^2 + (F2v - F2V)^2 + (F3v - F3V)^2}$$

Où $F1v$ représente la valeur de F1 en Bark d'une voyelle et $F1V$ la valeur moyenne en Bark de F1 pour la catégorie de cette voyelle (MOY_VOY). Par exemple, on soustrait à un exemplaire d'une voyelle /a/ la moyenne des valeurs de toutes les voyelles /a/ chez un même locuteur, ce qui donne la distance entre cet exemplaire et la moyenne de sa catégorie. La même formule a été appliquée pour F2 et F3. Cette mesure nous donne donc la distance euclidienne de chaque voyelle par rapport à son

[2] Par rapport à l'étude de (Huet & Harmegnies, 2000), nos DE_3D_VOY correspondent au $CM_{intra}$, tandis que nos DE_3D_SYS correspondent au $CM_{inter}$

centre de gravité sur trois dimensions. Nous pouvons ensuite calculer, pour chaque catégorie de voyelle la moyenne de ces DE. On obtient donc une valeur de dispersion pour chaque catégorie de voyelle de chaque locuteur.

### *2.2.2 Calcul des Distances Euclidiennes pour le système (DE_3D_SYS)*

DE_3D_SYS représente la dispersion des voyelles par rapport au centre de gravité du système (dispersion du système). Elle est obtenu en appliquant la formule suivante :

$$\sqrt{(F1v - F1S)^2 + (F2v - F2S)^2 + (F3v - F3S)^2}$$

Où $F1v$ représente la valeur de F1 en Bark d'une voyelle et $F1S$ la valeur moyenne en Bark de F1 pour l'ensemble des voyelles produites (MOY_SYS). La même formule a été appliquée pour F2 et F3. Cette mesure nous donne donc la distance euclidienne de chaque voyelle par rapport au centre de gravité de l'ensemble des voyelles produites (pour chaque locuteur) sur trois dimensions. Nous pouvons ensuite calculer, pour chaque locuteur la moyenne de ces DE. Toutefois, le calcul de cette moyenne est obtenu en regroupant les valeurs par catégorie de voyelle de façon à ne pas faire entrer dans le calcul le poids des effectifs de voyelles (Table 1). On obtient donc une valeur de dispersion du système vocalique pour chaque locuteur.

### *2.2.3 Calcul de l'Indice de Distinctivité (ID)*

L'Indice de Distinctivité (ID) a pour avantage de fournir une information dynamique sur l'espace de production des voyelles et de mettre en lien la dispersion du système avec celle des catégories de voyelle. Il est donc obtenu en calculant le rapport entre la dispersion du système et la dispersion de chaque catégorie de voyelle :

ID = moyenne DE_3D_SYS / moyenne DE_3D_VOY

# 3 Résultats

## 3.1 Centre de gravité (CG)

Nous avons, dans un premier temps, estimé le centre de gravité des productions vocaliques des locuteurs sur un plan F1/F2 (Figure 2). Ces CG sont calculés à partir des moyennes des catégories de voyelle et ne tiennent donc pas compte des effectifs de chaque catégorie de voyelle (Table 1).

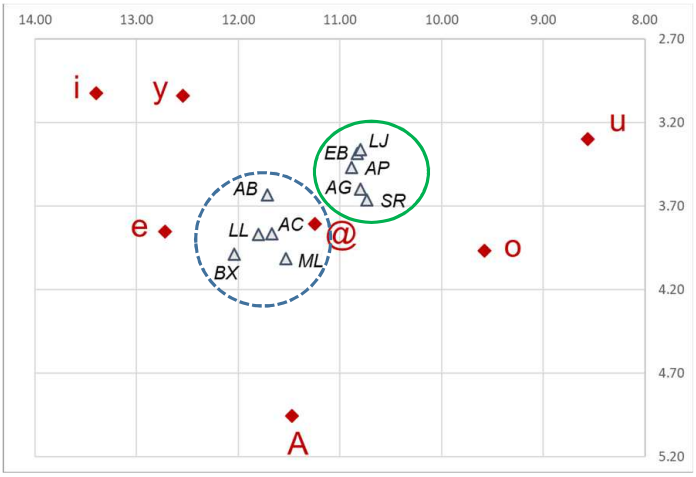



Figure 2 : Centre de gravité des productions des 10 locuteurs sur un plan F1/F2 (en Bark). Chaque locuteur est représenté par un triangle gris (hommes: cercle vert continu ; femmes: cercle bleu pointillé). En rouge sont représentées les valeurs moyennes de chaque catégorie de voyelle produites par l'ensemble des locuteurs.

Il en ressort une nette distinction homme/femme, les locutrices montrant un centre de gravité plus bas et plus antérieur tandis que celui des locuteurs est plus central et plus haut. Ce résultat est conforme aux observations habituelles et s'explique par la taille plus réduite du conduit vocal des locutrices. Globalement, le CG moyen de l'ensemble des locuteurs tend plutôt vers l'avant, ce qui n'est pas surprenant étant donné qu'une majorité des voyelles du français est plutôt antérieure (ou centrale) tandis qu'il n'y a que 2 voyelles d'arrière.

## 3.2 Distances euclidiennes 3D (F1-F2-F3)

Les DE_3D pour les voyelles et le système ont été calculées selon la méthode expliquée en 2.2.1 et 2.2.2. On note ainsi que les DE_3D_SYS sont toujours supérieures aux DE_3D_VOY (Table 2), ce qui est conforme à nos attentes puisque la dispersion autour de chaque voyelle est logiquement moins grande que la dispersion de toutes les voyelles par rapport au centre du système.

| | femmes | | | | | hommes | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AB | AC | BX | LL | ML | AG | AP | EB | LJ | SR |
| DE_3D_VOY | 1.39 | 1.78 | 1.53 | 1.58 | 1.51 | 1.63 | 1.73 | 1.63 | 1.40 | 1.50 |
| DE_3D_SYS | 1.94 | 2.22 | 2.07 | 2.11 | 2.01 | 2.20 | 2.23 | 1.97 | 1.94 | 2.12 |
| ID | 1.40 | 1.24 | 1.35 | 1.33 | 1.33 | 1.35 | 1.29 | 1.21 | 1.39 | 1.41 |

Table 2 : distance euclidienne 3D (F1, F2, F3) pour chacun des locuteurs et chacun des espaces (espace système et moyenne des espaces voyelles). L'ID correspond au rapport entre les deux DE_3D

On notera également un certain équilibre dans la production des locuteurs dans la mesure où, le plus souvent, lorsque la dispersion du système est faible, la dispersion de chaque voyelle tend à être également minimisée (AB, LJ) et inversement (AC, AP). Il semble donc que les locuteurs tendent vers un équilibre dans la production de leurs voyelles de façon à maintenir un minimum de distinctivité. Il est malgré tout possible de distinguer des profils différents en comparant les Indices de Distinctivité (ID).

## 3.3 Indices de distinctivité

Comme expliqué plus haut (2.2.3), l'ID se présente comme un rapport entre la dispersion du système et la dispersion des catégories de voyelles. La Figure 3 nous montre une représentation hiérarchisée des locuteurs les plus distinctifs (à gauche) vers ceux qui le sont le moins (à droite). Trois locuteurs (SR, AB et LJ) sont clairement au-dessus de la moyenne des locuteurs tandis que trois autres (AP, AC et EB) sont clairement en-dessous. Les quatre autres locuteurs se situent autour de la moyenne.

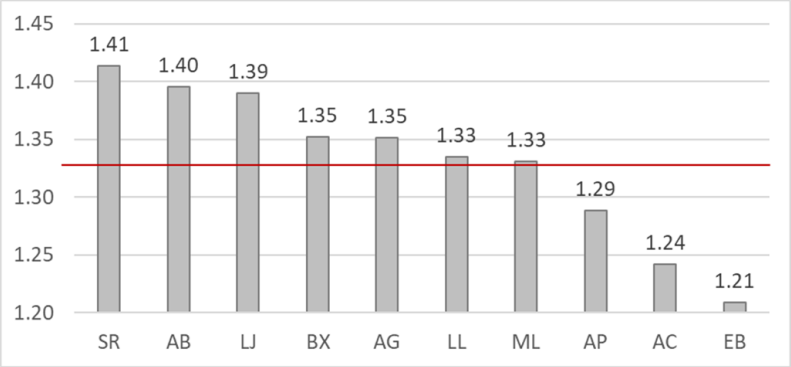



Figure 3: Indice de distinctivité (ID=rapport entre dispersion du système et dispersion des voyelles) calculé pour les 10 locuteurs. En rouge, la moyenne des ID des locuteurs (ID=1.33).

L'échelle laisse supposer que les différences sont importantes alors qu'en réalité, nous ne pouvons, en l'état, rien dire sur la magnitude de ces différences. Toutefois, nous pourrions poser l'hypothèse qu'une valeur 1 est une limite basse car en dessous de 1, cela signifie que la dispersion de chaque voyelle est plus forte que celle de tout le système, ce qui intuitivement est une limite. Cela aurait donc du sens de soustraire 1 à notre indice actuel de distinctivité, la valeur 0 indiquant alors une distinctivité nulle (dispersion de chaque voyelle étant égale à celle de tout le système). Dans ce cadre, une valeur à 1.41 (locuteur SR, Figure 3), qui deviendrait 0.41 après soustraction, indiquerait alors une nette différence avec 1.21 (locuteur EB) qui deviendrait 0.21. Dans tous les cas, pour pouvoir donner un sens à cette magnitude, il sera nécessaire de mettre en lien ces différences avec une mesure de l'intelligibilité des locuteurs.

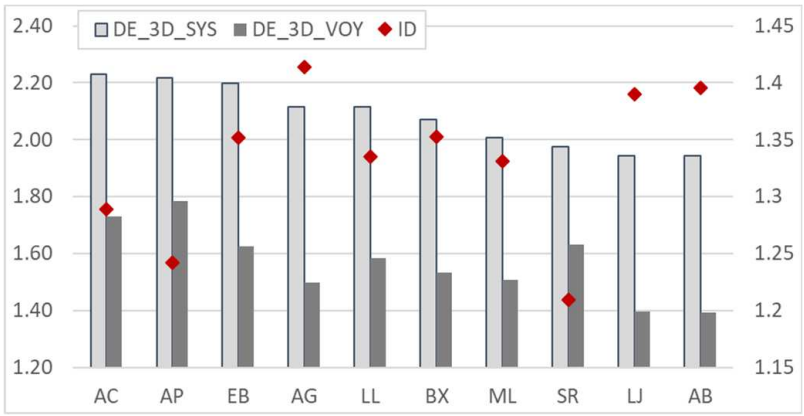



Figure 4: DE_3D_SYS (gris clair) et DE_3D_VOY (gris foncé) en Bark classés par ordre décroissant de DE_3D_SYS pour chaque locuteur (axe de gauche). En points rouges (axe de droite), l'ID pour chaque locuteur

Nous avons par ailleurs mis en lien l'ID avec les résultats concernant la dispersion du système, d'une part, et la dispersion des voyelles, d'autre part. Cette comparaison, représentée sur la Figure 4, met en évidence le fait qu'un ID élevé n'est pas forcément lié à un système dispersé. En effet, les locuteurs LJ et AB présentent des ID élevés mais les systèmes les moins dispersés. De même le locuteur SR a l'ID le plus élevé des locuteurs mais ne présente pas le système le plus dispersé. Ce qui explique l'ID élevé de ces trois locuteurs, c'est la faible dispersion des catégories de voyelle. Inversement, on notera que la locutrice AC, présentant un système fortement dispersé, a un ID faible en raison d'une forte dispersion des catégories de voyelle. Notre mesure de l'ID apparaît donc bien comme une combinaison des deux facteurs de dispersion et n'est pas systématiquement liée à la taille du système vocalique.

# 4 Conclusions et perspectives

Notre étude avait pour objectif de déterminer un indice de distinctivité basé sur l'indice *Phi* de Huet & Harmegnies (2000) et permettant de rendre compte de la relation dynamique entre la dispersion des catégories de voyelles et la dispersion de l'ensemble du système. Nous avons vu que cet indice permet de différencier les locuteurs. Par ailleurs, bien que les données sur les CG des systèmes des locuteurs montre une nette différence Homme/Femme, nous n'avons retrouvé cette différence ni dans les dispersions (système ou voyelles), ni dans la mesure de l'ID. Nous avons constaté également qu'un ID élevé n'est pas lié à un système plus large mais plutôt à la combinaison des deux facteurs.

Toutefois, nos conclusions doivent s'arrêter là et nous ne pouvons rien dire sur la magnitude de ces différences d'ID. En effet, rien ne nous permet de dire, à ce stade, que ces différences sont importantes ou qu'elles aient un impact sur l'intelligibilité des locuteurs. Pour cela, il nous faudra mettre en place une évaluation perceptive des productions de ces locuteurs. Notre objectif à plus long terme sera d'utiliser l'ID pour caractériser des contextes différenciés. En premier lieu, des travaux précédents ont pu mettre en évidence des espaces de réalisation distincts selon la langue parlée (Al-Tamimi & Ferragne, 2005; Meunier et al., 2006 ; Gendrot & Adda-Decker, 2007). Nous faisons l'hypothèse que, selon l'organisation du système vocalique mais également des propriétés lexicales présents dans une langue donnée, des profils de distinctivité différents pourraient émerger selon les langues. De même, on peut supposer que la magnitude des différences entre les ID des locuteurs sera plus importante si on observe leur production en lecture ou en parole spontanée (comme l'ont mis en évidence Huet & Harmegnies, 2000). On pourra alors observer si la hiérarchie du classement des locuteurs selon l'ID est conservée quelle que soit la situation de parole, ce qui supposerait des profils spécifiques aux locuteurs. Des expériences de type bite-block pourraient également nous permettre de mettre en évidence les phénomènes de compensation et de réajustement dynamique des dispersions. Enfin, les pathologies de la parole caractérisées par un déficit moteur (dysarthrie) montrent une désorganisation et/ou une centralisation du système vocalique pour lesquelles des systèmes de mesure très fins ont été proposés (Audibert & Fougeron, 2012). Nous pensons que la mesure de l'Indice de Distinctivité pourrait apporter des réponses complémentaires sur cette désorganisation qui n'est pas systématiquement une réduction du système vocalique. Cet indice de distinctivité pourrait finalement fournir une mesure de prédiction d'intelligibilité sachant qu'une faible distinctivité pourrait entrainer des problèmes de décodage et de compréhension.

## Remerciements

Les données audiovisuelles ont été enregistrées et mises à disposition dans le cadre du Centre d'Expérimentation sur la Parole du Laboratoire Parole et Langage à Aix-en-Provence.

## Références

AL-TAMIMI J. E., FERRAGNE E. (2005). Does vowel space size depend on language vowel inventories? Evidence from two Arabic dialects and French. *Proceedings of Interspeech*, 2464–1468.

AUDIBERT N., FOUGERON C. (2012). Distorsions de l'espace vocalique : quelles mesures? Application à la dysarthrie. *Proceedings of JEP-TALN-RECITAL,* 217–224.

BERTRAND R., BLACHE P., ESPESSER R., FERRE G., MEUNIER C., PRIEGO-VALVERDE B., & RAUZY S. (2008). Le CID — Corpus of Interactional Data — Annotation et Exploitation Multimodale de Parole Conversationnelle. *Traitement Automatique Des Langues*, *49* (3), 105–134.

GENDROT C., ADDA-DECKER M. (2005). Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German. *Proceedings of Interspeech*, 2453-2456.

GENDROT C., & ADDA-DECKER M. (2007). Impact of duration and vowel inventory size on formant values of oral vowels: an automated formant analysis from eight languages. *Proceedings of the 16th International Congress of Phonetic Sciences*, 1417–1420.

HUET K., & HARMEGNIES B. (2000). Contribution à la quantitication du degré d'organisation des systèmes vocaliques. *Journées d'Etudes sur la Parole,* 225–228.

MEUNIER C., ESPESSER R., & FRENCK-MESTRE C. (2006). Aspects phonologique et dynamique de la distinctivité au sein des systèmes vocaliques: une étude inter-langue. In *Journées d'Etude sur la Parole*, 333-336.

SMILJANIC R., BRADLOW A. R. (2009). Speaking and Hearing Clearly: Talker and Listener Factors in Speaking Style Changes. *Language and Linguistics Compass*, *3*(1), 236–264.

TRAUNMÜLLER, H. (1990). "Analytical expressions for the tonotopic sensory scale". *The Journal of the Acoustical Society of America*. 88: 97.

XXXIIe Journées d'Études sur la Parole
4-8 juin 2018, Aix-en-Provence, France

# Clarification et correction d'indices segmentaux : une étude pilote sur les consonnes occlusives du français

Maëva Garnier, Marion Dohen, Louis Buttiaux, Silvain Gerber
Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-lab, 38000 Grenoble, France
* Institute of Engineering Univ. Grenoble Alpes
maeva.garnier@gipsa-lab.grenoble-inp.fr

## RESUME

Cette étude aborde la question du trait de voisement et de la place d'articulation des consonnes occlusives du Français à l'aide d'un nouveau paradigme de correction segmentale, de façon à tester 1) quels indices de ces consonnes des locuteurs renforcent lorsqu'ils ont été mal compris par leur interlocuteur, et 2) si le renforcement de ces indices est toujours le même ou dépend de la consonne entendue à la place. Une locutrice jouait avec l'expérimentatrice à un jeu conçu pour simuler des situations d'incompréhension assez naturelles. La locutrice a montré plusieurs modifications de ses consonnes occlusives en parole claire et en situation de correction d'un mot mal compris (durée de la phase d'occlusion, VOT, intensité du bruit, …). En revanche, quasiment aucun descripteur n'a été modifié de façon différente en fonction du type d'erreur perceptive commise par l'interlocutrice (portant sur le trait de voisement ou sur le lieu d'articulation).

## ABSTRACT

**Clarification and correction of segment cues: a pilot study on French stop consonants.**
This study deals with the question of the voicing feature and the articulation place of French stop consonants, using a new paradigm of segment correction. It aims at testing 1) which acoustic cues speakers enhance when they are misunderstood by an interlocutor, and 2) whether this enhancement is always the same or depends on the misperceived consonant. One speaker played with the experimenter a game designed to simulate natural situations of misunderstanding. The speakers showed several modifications of her stop consonants in clear speech and in situations of segment corrections (duration of the occlusion phase, VOT, burst intensity, …). However, hardly no descriptor was modified in a different way, depending on the perceptual error made by the interlocutor (on the voicing feature or on the articulation place).

MOTS-CLÉS: Interaction face-à-face; Consonnes occlusives; Traits; Parole claire; Multimodalité
KEYWORDS: Face-to-face interaction; Stop consonants; Features; Clear speech; Multimodality

# 1 Introduction

Sur quels indices acoustiques nous basons-nous en tant que locuteur ou auditeur, pour distinguer les différentes catégories de consonnes occlusives (/p/, /b/, /t/, /d/, /k/, /g/ pour les consonnes occlusives orales du Français) ? Cette distinction est-elle particulièrement claire lorsque ces indices sont « prototypés », prenant des valeurs très spécifiques, ou bien lorsque ces indices sont « contrastés », prenant des valeurs les plus distinctes possibles entre deux catégories phonologiques ? Ces questions, loin d'être nouvelles, ont été traitées extensivement ces 50 dernières années par de nombreuses études phonétiques en production et perception de la parole.


10.21437/JEP.2018-55

Les études sur la production des consonnes occlusives ont décrit des différences physiologiques et acoustiques entre différentes catégories de consonnes. En Français, l'opposition entre des consonnes occlusives voisées et non voisées se base littéralement sur la présence vs. l'absence de vibration des plis vocaux, et par conséquent sur la présence vs. l'absence d'énergie périodique audible pendant la phase d'occlusion. En Français, les segments [b d g] montrent un VOT (Voice Onset Time) négatif tandis que les segments [p t k] sont caractérisés par un VOT positif. Cette principale différence de production a plusieurs autres conséquences, directes ou directes, sur la longueur de la transition du premier formant (F1) (Liberman et al. 1954), la fréquence initiale de sa réapparition (Lisker et al. 1978), la fréquence fondamentale (f0) à la reprise de voisement, après le relâchement de l'occlusion (Ohde 1984), la durée de la voyelle précédente (Delattre 1962; Abdelli-Beruh 2004) ou encore la durée de la phase d'occlusion (Ohala et Riordan 1979; Abdelli-Beruh 2004).
Le lieu d'articulation d'une consonne occlusive configure le volume des cavités avant et arrière du conduit oral. La cavité avant joue un rôle déterminant sur la fréquence de la 2ème et 3ème résonance du conduit vocal au moment du relâchement de l'occlusion et sur leur variation jusqu'à la voyelle suivante. Elle détermine également l'allure de l'enveloppe spectrale du bruit créé au relâchement de l'occlusion. Pour les occlusives labiales [p b], l'intensité du bruit de plosion est faible, avec un spectre diffus-descendant (Lousada et al. 2012; Forrest et al. 1988) et une transition du deuxième formant vocalique (F2) montante devant la plupart des voyelles. Pour les occlusives apico-alvéolaires (ou dentales, par abus de langage) [t d], la cavité avant du conduit vocal est étroite, contribuant à un bruit de plosion d'intensité moyenne, avec un spectre diffus-montant et une transition de F2 descendante devant des voyelles centrales et postérieures, et légèrement montante devant des voyelles antérieures. Enfin pour les occlusives vélaires (ou palatales, par abus de langage) [k g], l'occlusion fait converger le 2ème et le 3ème formant vocalique (F2 et F3), se traduisant par un bruit de plosion de forte intensité, au spectre compact. Stevens et Blumstein (1978) ont soutenu l'idée que ces caractéristiques spectrales soient relativement invariantes et spécifiques à chaque lieu d'articulation. Ces principales différences spectrales sont également accompagnées d'autres variations acoustiques : Plus l'occlusion du conduit vocal est postérieure, et plus le VOT est long (Cho et Ladefoged 1999 ; Lisker et Abramson 1964), la transition de F1 courte et sa fréquence de réapparition haute (Summerfield et Haggard 1977).
De nombreuses études perceptives ont confirmé ces contrastes observés en production, en montrant dans quelle mesure la variation isolée ou combinée de ces différents traits acoustiques affecte la catégorisation perceptive d'un son plosif. Ainsi, il a été confirmé que la perception du trait de voisement était en effet affectée par la variation du VOT (Serniclaes 1984; Williams 1977), de l'intensité du bruit de plosion (Repp 1978; 1983; Williams 1977), de la valeur initiale de f0 (Haggard 1981) et de F1 après le relâchement de l'occlusion (Liberman et al.,1954 ; Summerfield et Haggard 1977), par la longueur de la transition de F1 (Stevens et Klatt 1974), la durée de la phase d'occlusion et de la voyelle précédente (Lisker 1957; Crystal et House 1988). De même, il a été confirmé que la perception du trait de place d'articulation était en effet affectée par les variations spectrales du bruit de plosion (Gravel 1983), la direction des transitions du 2ème et du 3ème formant (Delattre et al. 1955; Harris et al. 1958), la longueur du VOT et des transitions formantiques (Lisker et Abramson 1970; Lisker 1975; Miller 1981).
Il reste néanmoins un certain nombre de questions concernant ces indices, en particulier la question de leur hiérarchie et de leur degré d'importance, possiblement variable en fonction des situations de communication, en particulier dans des situations où les indices principaux sont ambigus ou altérés (parole chuchotée, environnement bruyant) (Winn et al. 2013). C'est pourquoi nous proposons dans ce projet une nouvelle approche complémentaire de ces questions en examinant, au travers d'une expérience de production de parole en interaction face-à-face, 1) quels indices de ces consonnes des locuteurs renforcent lorsqu'ils ont été mal compris par leur interlocuteur, et 2) si le renforcement de ces indices est toujours le même ou dépend de la consonne entendue à la place.

# 2 Matériel et méthodes

Une locutrice de 24 ans, de langue maternelle française, sans trouble auditif ni de la parole a participé à cette expérience. Cette locutrice était naïve vis-à-vis de l'objet de l'expérience et ne connaissait aucun des expérimentateurs.

L'expérience consistait en un jeu interactif, dans lequel la locutrice devait donner des consignes à une interlocutrice (auteure MD) pour avancer dans un labyrinthe. Chacune des deux interlocutrices disposait devant elle, sur un écran, d'une grille de 5x5 cases sur laquelle étaient représentées la case de départ (verte) et la case d'arrivée (rouge) d'un chemin, ainsi que des pseudo-mots sur les autres cases (cf. Figure 1). Sur la grille de la locutrice étaient également représentés les murs d'un labyrinthe, déterminant le chemin pour aller de la case de départ à la case d'arrivée. La tâche de la locutrice consistait à décrire pas à pas à l'expérimentatrice les différentes étapes du chemin (déplacements d'une case en horizontal ou en vertical). L'expérimentatrice devait tracer sur la grille une flèche représentant le déplacement compris, tout en continuant de parler naturellement.
Le but du labyrinthe était de susciter des situations assez naturelles d'incompréhension. A chaque étape, le déplacement pouvait en effet aboutir à des cases présentant des mots très proches auditivement. De ce fait, l'expérimentatrice pouvait parfois faire semblant, de façon assez crédible et naturelle, de ne pas avoir compris la consigne. Un exemple d'interaction est donné ci-dessous.

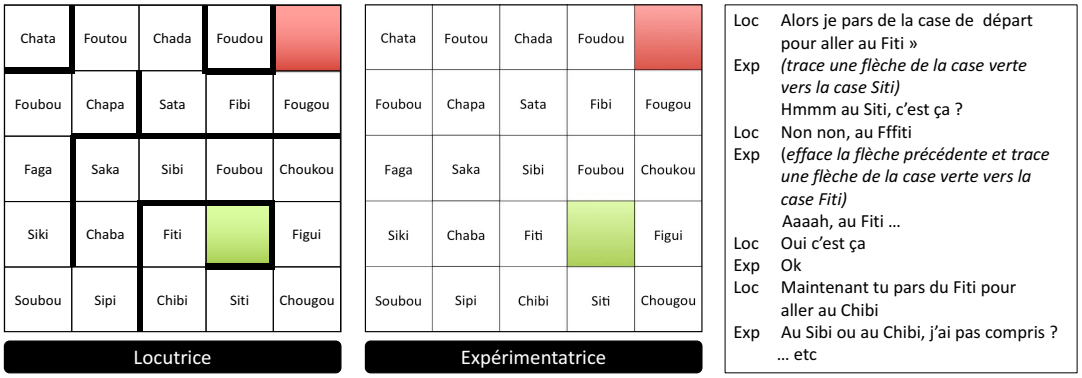



FIGURE 1: Exemple de grilles de jeu dont disposait la locutrice (à gauche) et l'expérimentatrice (à droite). La locutrice devait donner des consignes à l'expérimentatrice pour tracer le chemin partant de l'entrée du labyrinthe (case verte) à sa sortie (case rouge).

| | | C2 | | | | | |
|---|---|---|---|---|---|---|---|
| | | p | t | k | b | d | g |
| V | a | ʃapa | ʃata | ʃaka | ʃaba | ʃada | ʃaga |
| | i | sipi | siti | siki | sibi | sidi | sigi |
| | u | fupu | futu | fuku | fubu | fudu | fugu |

TABLE 1 : Liste des 18 mots-cibles de type $C_1VC_2V$ étudiés dans cette expérience.

Le jeu interactif permettait de faire produire à notre locutrice 18 mots cibles de type $C_1VC_2V$ où $C_2$ est une consonne occlusive du Français (/p/, /t/, /k/, /b/, /d/ ou /g/) en contexte vocalique /a/, /i/ ou /u/ (cf. Table1). Ces mots ont été choisis de façon à avoir des ensembles de 4 mots différant par un seul segment (ici une consonne occlusive), présentant soit le même trait de voisement, mais un lieu d'articulation différent (ʃapa vs. ʃata et ʃaka), soit le même lieu d'articulation mais une opposition de voisement (ʃapa vs. ʃaba). De tels quadruplés étaient difficilement trouvables dans le lexique Français. C'est pourquoi nous avons fait le choix de pseudo-mots, traités comme des noms propres dans le contexte du jeu. Aucun d'entre eux ne correspondait à un mot réel, si bien que nous pouvons considérer leur familiarité comme identique. Pour le naturel du jeu, le corpus comportait également

36 autres pseudo-mots se distinguant de nos 18 mots cibles par la consonne initiale (C1) ou par la voyelle (V). Ces mots n'étaient que des « fillers » et n'ont pas été analysés.

Les deux interlocutrices interagissaient dans deux pièces séparées, à l'aide d'un système de visioconférence. Les deux interlocutrices étaient assises et portaient un micro-casque (AKG HSD 171), relié à une carte son (RME Fireface 800) permettant en même temps d'acquérir les deux signaux audio (à fe=44.1 kHz) et de renvoyer dans le casque de chaque participante un retour calibré de sa propre voix et celle de sa partenaire. La locutrice était filmée avec une caméra (25 images/s) située face à elle, derrière un écran/prompteur sur lequel était diffusé la vidéo de l'expérimentatrice. L'expérimentatrice était également filmée de face, grâce à une webcam miniature positionnée au centre de l'écran placé devant elle, diffusant la vidéo de la locutrice. Ce dispositif permettait que les interlocutrices interagissent en se regardant dans les yeux. En plus de cet écran principal, la locutrice disposait également d'un écran sur la gauche, représentant sa grille et d'un écran sur la droite, représentant la grille de l'expérimentatrice, et diffusant en temps réel les flèches tracées par l'expérimentatrice sur une tablette graphique.

L'expérience comportait 23 parties du jeu, avec une grille différente dans chaque partie. Durant les 5 premières parties, les deux interlocutrices interagissaient en condition de visioconférence « normale », sans perturbation (condition SP). Ces parties se déroulaient facilement, sans aucune incompréhension, et servaient de référence à la production de parole conversationnelle par la locutrice. Les 5 parties permettaient d'enregistrer 3 occurrences de chacun des 18 mots-cibles en parole conversationnelle.

Durant les 18 parties suivantes, la locutrice était informée qu'une perturbation était introduite sur le canal audio du système de visioconférence, de telle façon que l'expérimentatrice allait avoir des difficultés à la comprendre. En réalité, aucune perturbation n'était introduite mais l'expérimentatrice faisait effectivement semblant de mal entendre, amenant la locutrice à parler de façon plus claire (condition APN). L'expérimentatrice disposait d'un scénario très contrôlé lui indiquant, étape par étape de chaque grille, ce qu'elle devait faire semblant d'avoir compris. Ainsi, assez régulièrement (mais non systématiquement), l'expérimentatrice faisait semblant de se tromper, poussant la locutrice à répéter sa consigne, et donc le mot cible, en le corrigeant par rapport au mot mal compris par l'expérimentatrice (condition APC) (cf. exemple d'interaction Figure 1). De façon contrôlée, l'incompréhension (et donc la correction qui en découlait) portait soit sur le trait de voisement (1/3 du temps), soit sur le lieu d'articulation (1/3 du temps pour chacun des 2 autres lieux d'articulation que celui de la consonne cible, par exemple lieu dental ou palatal pour une consonne labiale, et réciproquement). Les 18 parties permettaient ainsi d'enregistrer 6 occurrences de chacun des 18 mots-cibles en parole claire, suivies de leur correction suite à une incompréhension simulée par l'expérimentatrice (réparties en 2 occurrences pour les 3 types d'erreur : sur le voisement, et sur les 2 autres lieux d'articulation).

Les données ont été étiquetées manuellement sous Praat en repérant le début et la fin des mots-cibles (t0 et t7), l'instant de disparition du 2ème formant (F2) à la fin de la voyelle précédente (t1) et celui de réapparition du F2 au début de la voyelle suivante (t6), pour les consonnes non voisées, l'instant de disparition du voisement après la fin de la voyelle précédente (t2) et celui de réapparition du voisement avec ou après le bruit de plosion (t5), enfin les instants de début (t3) et de fin (t4) du bruit (plosion+friction) émis au relâchement de l'occlusion.

A l'aide de scripts développés sous Matlab, différents descripteurs ont ensuite été extraits du signal audio sur ces différents intervalles de temps : la durée de la consonne (t7-t1), de la phase d'occlusion (t3-t1) et du bruit (t4-t3), le VOT, tel que défini par Lisker (t5-t3), le pourcentage de temps durant lequel les plis vocaux continuaient de vibrer pendant la phase d'occlusion pour les consonnes non voisées (taux de voisement : (t2-t1)/(t3-t1), l'intensité acoustique moyenne du voisement, s'il y en a, pendant la phase d'occlusion, et l'intensité du bruit.

Les analyses statistiques ont été réalisées avec le logiciel R, de façon séparée pour tester :

l-l'effet de parler plus clairement en situation de communication perturbée. Pour chaque descripteur, nous avons modélisé les données à l'aide d'un modèle linéaire incluant le facteur Condition (2 niveaux : APN et SP), le facteur Consonne (6 niveaux : /p/, /t/, /k/, /b/, /d/, /g/) et le facteur Voyelle (3 niveaux : /a/, /i/, /u/). Les mots n'étaient pas appariés entre les conditions SP et APN.

2-l'effet de corriger un mot suite à une incompréhension de l'interlocutrice. Pour tenir compte de l'appariement entre la première occurrence d'un mot et sa deuxième occurrence répétée, nous avons choisi de considérer comme variables dépendantes la variation de nos descripteurs (Δ) entre les conditions APN et APC. Pour chacune de ces variations de descripteur, nous avons modélisé les données à l'aide d'un modèle linéaire incluant le facteur Voyelle (3 niveaux : /a/, /i/, /u/) et le facteur Type d'erreur (18 niveaux : consonne voisée prise pour une non voisée, non voisée prise pour une voisée, labiale prise pour une dentale, labiale prise pour une vélaire, etc…).

Suivant la même procédure que celle développée avec des experts statisticiens, appliquée à l'analyse de plusieurs jeux de données précédents (Bourne et al. 2016), nous avons commencé par simplifier chaque modèle en excluant toute interaction non significative entre les facteurs. Une fois le modèle simplifié, nous avons vérifié sa validité en examinant ses résidus. Nous avons ensuite réalisé des tests de modèle emboités (ou Likelihood Ratio Test) pour tester la significativité de chaque facteur ou de leurs interactions. Enfin, nous avons réalisé des tests post-hoc pour examiner le contraste plus spécifique entre certaines conditions, en appliquant des corrections de Bonferroni pour les comparaisons multiples.

# 3 Résultats

Les analyses statistiques ont montré que la locutrice allongeait significativement la durée de ses consonnes occlusives lorsqu'elle parlait plus clairement en situation perturbée (APN vs. SP : Δ=+41ms, $p<0.0001$), avec un allongement semblable quels que soient la consonne occlusive et le contexte vocalique (modèle DuréeConsonne ~ Condition + Voyelle + Consonne). La durée des consonnes était encore plus allongée lors d'une correction, suite à une incompréhension de l'interlocutrice (APC vs. APN : Δ=+66ms, $p<0.0001$), de nouveau sans effet significatif du contexte vocalique, de la consonne ou du type d'erreur sur cet allongement (modèle ΔDuréeConsonne ~ 1).

Une analyse temporelle plus détaillée a montré que cet allongement de la consonne provenait essentiellement d'un allongement significatif de la phase d'occlusion en parole claire (APN vs. SP : Δ=+44ms, $p<0.0001$ (modèle DuréeOcclusion ~ Condition + Consonne)) et lors d'une correction (APC vs. APN : Δ=+52ms en moyenne pour les consonnes en contexte /a/ et /i/, $p<0.0001$ ; Δ=+93ms en moyenne pour les consonnes en contexte /u/, $p<0.0001$ (modèle ΔDuréeConsonne ~ Voyelle)).

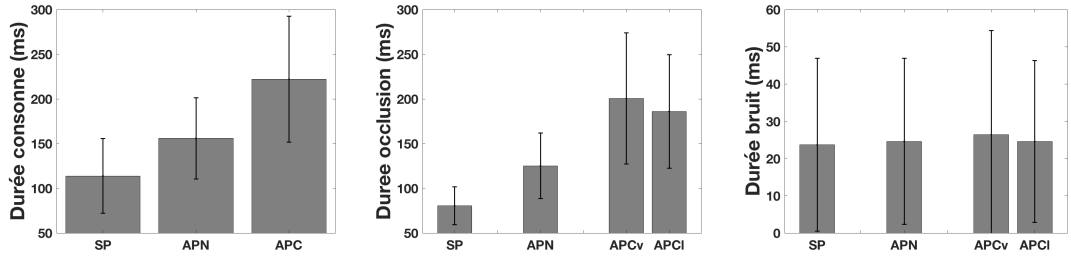



FIGURE 2: Evolution de la durée des consonnes occlusives, de leur phase d'occlusion et de leur bruit de plosion, lorsque la locutrice parlait de façon conversationnelle, dans une condition non perturbée (SP), lorsqu'elle parlait plus clairement en situation perturbée (APN), et lorsqu'elle corrigeait une erreur perceptive de son interlocutrice, commise sur le trait de voisement (APCv) ou sur le trait d'articulation (APCl).

Lors de la correction, l'allongement n'était en tout cas pas significativement affecté par le type d'erreur commise par l'auditeur (erreur sur le trait de voisement vs. sur le lieu d'articulation).
L'allongement de la consonne n'avait en revanche aucun effet significatif sur la durée du bruit de plosion, celui-ci ne variant significativement ni en parole claire, ni lors d'une correction.

Les consonnes occlusives voisées montrent une vibration des plis vocaux tout le temps de cette phase d'occlusion, tandis que les occlusives non voisées ne montrent pas de vibration, ou bien durant une courte durée après la voyelle précédente. Cette énergie basse fréquence « résiduelle » pourrait entraîner une certaine confusion perceptive, donnant à l'auditeur l'impression de voisement. C'est pourquoi nous nous attendions à ce que la durée de ce voisement résiduel et son intensité acoustique, soit diminuées lorsqu'un locuteur cherche à améliorer son intelligibilité.
En accord avec nos prédictions, notre locutrice tendait à diminuer la durée de ce voisement résiduel durant la phase d'occlusion pour ses consonnes non voisées, de façon non significative lorsqu'elle parlait plus clairement (APN vs. SP : $\Delta$= -4.6% (modèle TauxVoisementOcclusion ~ Voyelle + Consonne)) mais de façon plus marquée lors d'une correction (APC vs. APN : $\Delta$= -5.5%, p=0.0001), sans pour autant que cette réduction soit significativement plus marquée lorsque l'erreur commise par l'interlocutrice concernait le trait de voisement, par rapport à une erreur sur le lieu d'articulation (modèle $\Delta$TauxVoisementOcclusion ~1).
En revanche, contrairement à nos prédictions, la locutrice augmentait globalement l'intensité moyenne du voisement pendant la phase d'occlusion lorsqu'elle parlait plus clairement, qu'il s'agisse du voisement « normal » des consonnes voisées ou du voisement résiduel des non voisées (APN vs. SP : $\Delta$= +6.6 dB, p<0.001 (modèle IVoisementOcclusion ~ Condition + Voyelle + Consonne)). A l'inverse, lors d'une correction, la locutrice tendait plutôt à diminuer l'intensité du voisement pendant la phase d'occlusion, que ce soit pour les consonnes voisées ou non voisées (APC vs. APN : $\Delta$= - 1.7 dB, p<0.001), avec cependant une tendance plus marquée lorsqu'il s'agissait de consonnes non voisées ayant été incorrectement perçues comme voisées par l'interlocutrice (cf. Figure 3).

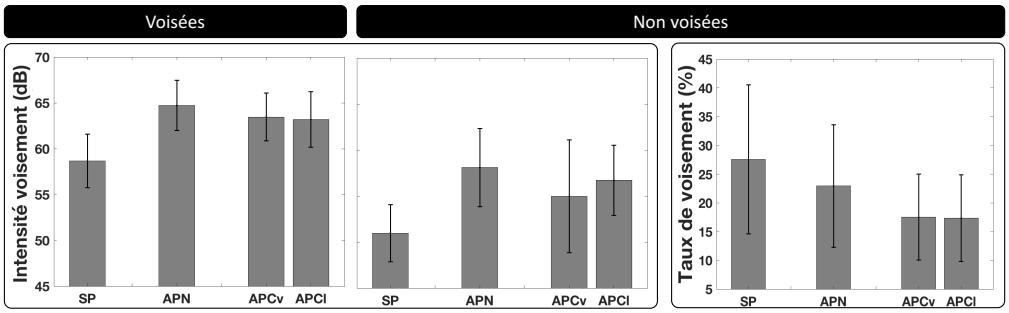



FIGURE 3: Evolution de l'intensité moyenne du voisement pendant la phase d'occlusion pour des consonnes occlusives voisées, ainsi que de la durée (relative) de voisement résiduel après la voyelle précédente durant la phase d'occlusion et de son intensité pour des consonnes occlusives non voisées, lorsque la locutrice parlait de façon conversationnelle, dans une condition non perturbée (SP), lorsqu'elle parlait plus clairement en situation perturbée (APN), et lorsqu'elle corrigeait une erreur perceptive de son interlocutrice, commise sur le trait de voisement (APCv) ou sur le trait d'articulation (APCl).

Le VOT, indice déterminant pour la perception du trait de voisement, est positif pour les consonnes occlusives du Français, plus long pour les consonnes vélaires [k] que les consonnes labiales et dentales [p t] (Lisker et Abramson, 1964). Contrairement à ce que nous aurions pu attendre, notre locutrice n'allongeait pas le VOT de toutes ses consonnes non voisées lorsqu'elle parlait plus clairement. Le VOT était significativement raccourci pour les consonnes /p/ et /t/ (APN vs. SP : $\Delta$=- 6ms, p<0.0001) et ne variait pas significativement pour les consonnes /k/ (modèle VOT ~ Condition*Consonne + Voyelle*Consonne). Lors d'une correction (APC vs. APN), le VOT était

légèrement allongé pour les consonnes non voisées en contexte /i/ (Δ=+8ms, p=0.014) et ne montrait pas de variation significative pour les autres contextes vocaliques (modèle ΔVOT ~ Voyelle). De façon intéressante, le VOT montrait plutôt une tendance à la diminution pour les consonnes /b/ (pour lesquelles il est déjà plutôt court) et à l'augmentation pour les consonnes /k/ (pour lesquelles il est déjà plutôt long). Sa variation n'était pas significativement affectée par le type d'erreur perceptive commise par l'interlocutrice, portant sur le trait de voisement ou sur le lieu d'articulation de la consonne.

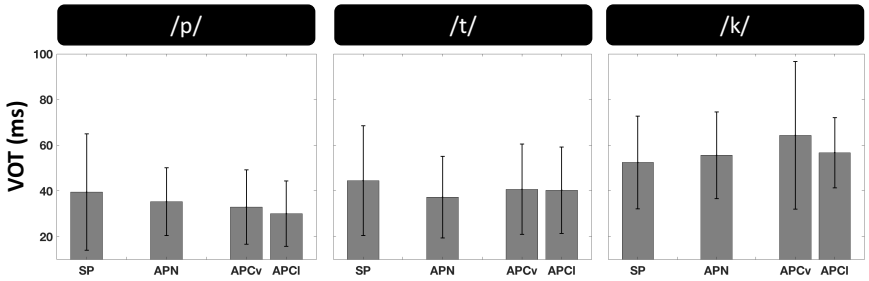



FIGURE 4: Evolution du VOT (délai d'établissement du voisement) des consonnes occlusives non voisées lorsque la locutrice parlait de façon conversationnelle, dans une condition non perturbée (SP), lorsqu'elle parlait plus clairement en situation perturbée (APN), et lorsqu'elle corrigeait une erreur perceptive de son interlocutrice, commise sur le trait de voisement (APCv) ou sur le trait d'articulation (APCl).

Les différences d'intensité et de spectre du bruit de plosion, enfin, renseignent sur le lieu d'articulation de la consonne (Stevens et Blumstein, 1978). Conformément à la littérature, nous avons effectivement bien observé chez notre locutrice des bruits de plosion assez faibles pour les consonnes occlusives labiales. En revanche, l'intensité des bruits de plosion s'est avérée relativement comparable pour ses consonnes occlusives dentales et vélaires. Dans tous les cas, notre locutrice augmentait significativement l'intensité du bruit de plosion de ses consonnes occlusives lorsqu'elle parlait plus clairement (APN vs. SP : Δ=+7.9 dB, p<0.0001), de façon comparable pour toutes les consonnes (modèle IBruitPlosion ~ Condition + Voyelle*Segment). Lors d'une correction (APC vs. APN), l'intensité du bruit de plosion était encore davantage renforcée pour des consonnes occlusives en contexte vocalique /u/ (+3.4 dB, p=0.005) mais ne variait pas significativement pour les autres contextes (modèle ΔIBruitPlosion ~ Voyelle). Les variations d'intensité du bruit de plosion ne dépendaient pas significativement de l'erreur perceptive commise par l'interlocutrice, en particulier du lieu d'articulation incorrectement perçu (labial, dental ou palatal).

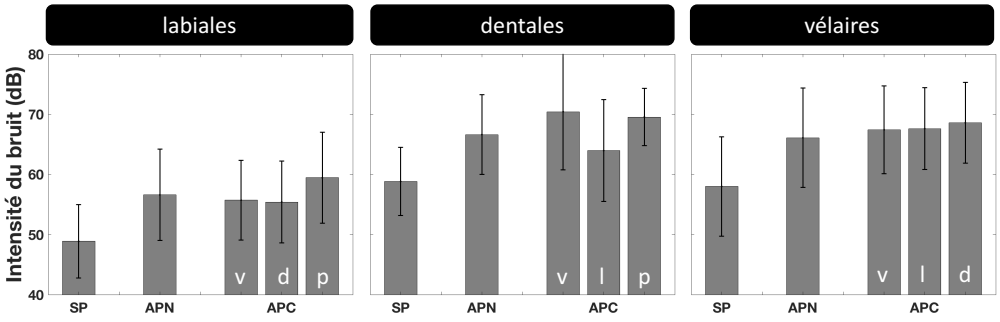



FIGURE 5: Evolution de l'intensité du bruit de plosion des consonnes occlusives labiales [p b], dentales [t d] ou vélaire [k g] lorsque la locutrice parlait de façon conversationnelle, dans une condition non perturbée (SP), lorsqu'elle parlait plus clairement en situation perturbée (APN), et lorsqu'elle corrigeait une erreur perceptive de son interlocutrice, commise sur le trait de voisement (APCv) ou sur les deux autres lieux d'articulation que celui de la consonne produite (APCl pour un lieu d'articulation incorrectement perçu comme labial, APCd pour dental, APCp pour palatal).

# 4 Discussion et conclusion

Sur quels indices acoustiques nous basons-nous en tant que locuteur ou auditeur, pour distinguer les différentes catégories de consonnes occlusives, et lesquels d'entre eux renforçons-nous lorsque nous devons améliorer notre intelligibilité en situation de communication perturbée ou lorsque notre interlocuteur nous a mal compris ?
Avec toutes les réserves liées au fait que nous avons pour l'instant exploré cette question sur une seule locutrice et sur un ensemble restreint de descripteurs acoustiques, nous pouvons néanmoins déjà dire que pour améliorer son intelligibilité, la locutrice de notre expérience pilote a montré plusieurs modifications significatives de la production de ses consonnes occlusives, allant dans le même sens en parole claire (APN vs. SP) et lors d'une correction (APC vs. APN), avec une modification plus marquée lors d'une correction: un allongement de la durée de ses consonnes occlusives provenant essentiellement d'un allongement de leur phase d'occlusion, une diminution de la durée du voisement résiduel à la fin de la voyelle précédente pendant la phase d'occlusion de ses consonnes non voisées, une diminution du VOT sur les consonnes /p/ et une augmentation sur les consonnes /k/, une augmentation de l'intensité du bruit de plosion-friction au relâchement de l'occlusive. Ces modifications acoustiques sont globalement conformes à nos attentes et peuvent s'interpréter en termes de stratégie communicationnelle visant à améliorer l'audibilité des indices et leur temps de récupération (intensité du bruit, durée de la phase d'occlusion), mais visant également possiblement à renforcer le contraste inter-catégoriel (VOT des consonnes non voisées labiales et vélaires, durée du voisement résiduel pendant la phase d'occlusion des consonnes non voisées).

Cette distinction des différentes catégories de consonnes occlusives est-elle particulièrement claire lorsque ces indices sont « prototypés », prenant des valeurs très spécifiques, ou bien lorsque ces indices sont « contrastés », prenant des valeurs les plus distinctes possibles entre deux catégories phonologiques ?
Contrairement à nos attentes, seul un des descripteurs examinés, la durée de voisement résiduel à la fin de la voyelle précédente pendant la phase d'occlusion des consonnes non voisées, a été davantage diminué par la locutrice suite à des erreurs perceptives de son interlocutrice commises sur le trait de voisement, par rapport à des erreurs commises sur le lieu d'articulation. La modification des autres descripteurs ne s'est pas montré affectée par le type d'erreur perceptive commise par l'interlocutrice, laissant penser que lors d'une correction, notre locutrice ne cherchait pas tant que cela à renforcer des contrastes acoustiques.

La prochaine étape de ce projet consistera, très naturellement, à étendre ces premières analyses à un ensemble plus complet de descripteurs acoustiques et articulatoires des consonnes occlusives (spectre du bruit, transitions formantiques, degré de compression des lèvres lors d'une occlusion labiale, etc) et à généraliser (ou non) ces observations à une cohorte de 10-15 locuteurs.

# Remerciements

Nous remercions Christophe Savariaux et Frédéric Elisei pour leur aide lors de la mise au point de l'expérience et de l'acquisition des données.
Cette recherche est financée par l'Agence Nationale de la Recherche (Projet StopNCo : Effort et coordination dans la production des consonnes occlusives ; ANR-14-CE30-0017; Maëva Garnier).

# Références


ABDELLI-BERUH. N. (2004). The Stop Voicing Contrast in French Sentences: Contextual Sensitivity of Vowel Duration, Closure Duration, Voice Onset Time, Stop Release and Closure Voicing, Journal: Phonetica, vol. 61, no. 4, pp. 201-219.

BOURNE, T., GARNIER M., SAMSON A. (2016). Physiological and acoustic characteristics of the male music theatre voice. The Journal of the Acoustical Society of America, 140(1), 610-621

CARLSON, R., GRANSTRÖM, B., PAULI, S. (1972). Perceptive evaluation of segmental cues. STL/QPSR 1/1972, Stockholm, Roy. Inst.Technol.; 18-24.

CHO, T., LADEFOGED, P. (1999). Variation and universals in VOT: evidence from 18 languages. Journal of Phonetics, 27(2), 207-229.

CRYSTAL T., HOUSE A. (1988). A note on the variability of timing control. Journal of Speech and Hearing Research, 31, 497-502.

DELATTRE P., LIBERMAN A., COOPER F. (1955). Formant transitions and loci as acoustic correlates of place of articulation in American fricative consonants. Studia Linguistica, 16, 104-121.

DELATTRE, P. (1962) Some factors of vowel duration and their cross-linguistic validity, JASA, 34, 1141-1142.

DORMAN M. F., STUDDERT-KENNEDY M., RAPHAEL L. (1977). Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. Perception and Psychophysics, 22, 109-122

FORREST, K., WEISMER, G., MILENKOVIC, P., DOUGALL, R. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. The Journal of the Acoustical Society of America, 84(1), 115-123.Stevens and Blumstein 1978

FRANCIS, A., KAGANOVICH, N., DRISCOLL-HUBER, C. (2008). Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. The Journal of the Acoustical Society of America, 124(2), 1234-1251.

GRAVEL, J., OHDE, R. (1983). Perception of stop place of articulation: Effects of stimulus amplitude. American Speech-Language-Hearing Association, 25, 101.

HARRIS, K.S. (1958). Cues for the discrimination of American English fricatives in spoken syllables. Language and Speech 1, 1-7.

LIBERMAN A., DELATTRE P., COOPER F., GERSTMAN L. J., (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychological Monographs, 68, 1-13.

LISKER, L. (1957). Closure duration and the intervocalic voiced- voiceless distinction in English. Language 33, 42-49.

LISKER, L., ABRAMSON, A. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. Word, 20(3), 384-422.

LISKER, L., ABRAMSON,A. (1970). The voicing dimension: some experiments in comparative phonetics. Proc. of the 6th Int. Cong. of Phonetic Sciences, Prague 1967; Prague: Academia, 1970; 563-567.

LISKER, L. (1975). Is it VOT or a first formant transition detector? Acoust.Soc.Am. 57, 1547-1551

LOUSADA, M., JESUS, L., PAPE, D. (2012). Estimation of stops' spectral place cues using multitaper techniques. DELTA 28(1), 1-26.

MILLER, J. (1981). Phonetic perception: Evidence for context- dependent and context-independent processing. J.Acoust.Soc.Am. 69, 822-831.

OHALA, J., RIORDAN,C. (1979). Passive vocal tract enlargement during voiced stops" in Speech Comunication Papers, J. Wolf and D. Klatt eds.; Acoust.Soc.Am.: New York; 89-92.

OHDE, R. (1984). Fundamental frequency as an acoustic correlate of stop consonant voicing. The Journal of the Acoustical Society of America, 75(1), 224-230.

POLS, L., SCHOUTEN, M. (1985). Plosive consonant identification in ambiguous sentences. J.Acoust.Soc.Am. 78, 33-39.

REPP, B., LIBERMAN, A., ECCARDT, T., PESETSKY, D. (1978). Perceptual integration of acoustic cues for stop, fricative, and affricate manner. Journal of Experimental Psychology: Human Perception and Performance, 4(4), 621.

SERNICLAES, W. (1984). Fenêtre de prélèvement temporel des indices d'occlusives. Dans Actes des XXIèmes Journées d'Etudes sur la Parole, 67-78.

STEVENS, K., KLATT, D. (1974). Current models of sound sources for speech. In Ventilatory and phonatory control systems: and international symposium. Oxford University Press, New York.

STEVENS, K., MANUEL, S., MATTHIES, M. (1999). Revisiting place of articulation measures for stop consonants: Implications for models of consonant production. In Proceedings of the International Congress of Phonetic Sciences (pp. 1117-1120).

STUDDERT-KENNEDY M. (1990). Language development from an evolutionary perspective. Haskins Laboratories Status Report on Speech Research, 101-102, 14-27.

SUMMERFIELD, Q., HAGGARD, M. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. The Journal of the Acoustical Society of America, 62(2), 435-448.

WILLIAMS, E. (1977). Experimental comparisons of face-to-face and mediated communication: A review. Psychological Bulletin, 84(5), 963.

WINN, M., CHATTERJEE, M., IDSARDI, W. (2013). Roles of Voice Onset Time and F0 in Stop Consonant Voicing Perception: Effects of Masking Noise and Low-Pass Filtering. Journal of Speech, Language, and Hearing Research, 56(4), 1097-1107.

# Déficit phonético-phonologique dans l'aphasie

Typhanie Prince

UMR1253, iBrain, équipe 1 Psy. Neurofonctionelle, INSERM, Université de Tours, 3 rue des Tanneurs, 37041, Tours, Cedex 1, France
Laboratoire de Linguistique, LLING, UMR 6310, Nantes
Typhanie.prince@gmail.com

## RÉSUMÉ

Cette étude présente une analyse linguistique des troubles phonético-phonologiques dans l'aphasie de Broca et de Wernicke. À partir de l'observation des productions de 13 locuteurs aphasiques francophones, lors de la réalisation de 40 items contenant des séquences consonantiques complexes (*via* une tâche de dénomination), nous revenons sur la question de l'origine des troubles à travers la dichotomie phonétique – phonologie et en offrons un approfondissement. Si la nature de l'aphasie engendre différents *patterns* d'erreurs, certains contextes phonologiques sont particulièrement affectés. Les paraphasies relevées chez ces locuteurs sont loin d'être aléatoires. Celles-ci se réduisent à cinq types : substitution, omission, épenthèse, métathèse et réduction. L'ensemble des résultats révèle une interaction entre segments, syllabes et position et questionne tant la phonologie que la neuropsycholinguistique. Pour comprendre la nature des transformations générées dans l'aphasie et montrer le lien entre troubles phonétiques et phonologiques, nous proposons une analyse des substitutions dans le cadre de la théorie des éléments.

## ABSTRACT

**Phonetic and phonological impairments in Aphasia**
This paper presents a linguistic analysis of the phonetical and phonological impairments in 13 French aphasic speakers suffering from Broca's and Wernicke's aphasia. From a study based on the production of 40 items containing different consonantal sequences through a picture-naming task, we discuss the phonetic *versus* phonological's origin of the impairments in order to propose a new reflexion. The detected paraphasias applied during a phonological deficit are not random. They are reduced to five kinds: substitution, deletion, epenthesis, metathesis and total reduction. These results reveal an interaction between segments, syllables and position and open the questions for both formal phonology and neuropsycholinguistics. Certain syllabic positions and contexts are prone to particular types of paraphasias. In order to understand the nature of the transformations generated during an aphasia, and to demonstrate the relation between phonetical and phonological deficits, we offer an analysis of the consonantal substitutions in Element Theory framework.

MOTS-CLÉS : Aphasies de Broca et Wernicke, phonologie, substitutions, théorie des éléments.
KEYWORDS: Brocas' aphasia, Wernicke's aphasia, phonology, substitutions, element theory.


10.21437/JEP.2018-56

# 1 Introduction

L'aphasie constitue une fenêtre ouverte sur les processus linguistiques mentaux. Elle est une source « d'évidences externes » (Nespoulous, 2006) et permet de valider ou de récuser les hypothèses formulées pour décrire le fonctionnement du langage (Forest, 2005). Lorsqu'elle est de type vasculaire – mais pas uniquement –, elle entraîne fréquemment un déficit de nature (lexico) phonético-phonologique. Trois types d'atteintes peuvent engendrer ce trouble : (1). une atteinte articulatoire, c'est-à-dire un trouble de la planification, de la programmation motrice des unités sur un plan phonétique, (2). un accès erroné aux représentations phonologiques sous-jacentes, c'est-à-dire une déficience dans l'accès aux informations contenues dans le système phonologique, et encore (3), une affection des représentations phonologiques elles-mêmes. Quelle que soit la cause, les structures syllabiques et segmentales sont affectées lors des réalisations orales. De manière générale, les troubles phonétiques, qui résultent d'un déficit moteur et articulatoire, c'est-à-dire de la planification, sont principalement attribués à l'aphasie de Broca (ainsi qu'à l'anarthrie pure où ces derniers s'apparentent au syndrome de désintégration phonétique, Alajouanine, Ombredane, Durand, 1939). Les troubles phonologiques touchant la sélection et l'ordonnancement des phonèmes sont, quant à eux, plus souvent attribués à l'aphasie de Wernicke ou de conduction. L'ensemble de ces déficits reflètent donc des processus physiopathologiques différents avec, d'une part, une difficulté à planifier la coordination des mouvements articulatoires, et, d'autre part, une sélection de phonèmes inappropriée au niveau lexical. Or, il est fréquent de rencontrer des locuteurs aphasiques de Broca souffrant également d'un trouble phonologique et vice versa. En effet, même si l'origine clinique du trouble et le lieu de la lésion ne sont pas semblables, il est souvent difficile d'établir si l'origine des transformations réalisées sont la conséquence d'un déficit purement phonétique ou/et phonologique. Autrement dit, la frontière entre ces deux plans – phonétique et phonologique – est loin d'être précisément tracée.

Nous cherchons dans cet article à comprendre si les locuteurs aphasiques de Broca, atteint d'un déficit dit *phonétique*, réalisent, eux aussi, des transformations directement imputées à la structure phonologique, comme celles produites par les locuteurs aphasiques de Wernicke et de conduction. À l'instar de Buckingham et Christman (2008 :127), nous appuyons l'idée d'une interface des deux niveaux où le fonctionnement est connexe : « there is evidence that there are no clear cut linguistic and neurological divisions in these phonetic/phonemic systems and this in turn forces aphasiologists to consider their interactions in describing phonetic and phonological breakdowns subsequent to brain damage ».

Les séquences consonantiques du français et leurs réalisations par des locuteurs aphasiques constituent des observatoires privilégiés. En effet, si la syllabe forme une unité particulièrement affectée, et que sa structure engendre différents types d'opérations phonologiques en surface, l'observation de ces transformations peut nous renseigner sur le fonctionnement des mécanismes sous-jacents. Nous observons ici les séquences consonantiques complexes du français ([s]+occlusive et [ʁ]+occlusive, désormais sT et RT) et leurs réalisations orales dans deux tableaux cliniques : Broca, avec et sans apraxie de la parole, et Wernicke. Cette étude se veut essentiellement phonologique et qualitative, car c'est à travers l'analyse fine de certaines opérations que nous observons des indices sur l'origine du trouble[1].

Dans la première section, nous brosserons un portrait des troubles phonologiques dans l'aphasie *via* l'analyse des paraphasies phonologiques dans les études récentes. Une seconde section offrira, à partir de nos données, un panorama des transformations consonantiques chez les deux populations et montrera comment ces dernières renseignent la théorie phonologique. Nous

[1] Une étude du signal acoustique sera réalisée dans un travail de recherche à venir.

montrerons que les transformations réalisées ne relèvent pas du hasard et, qu'indépendamment de la variation et de la nature de la lésion, toutes respectent les contraintes phonologiques de la langue maternelle et semblent suivre un schéma qui répond à la notion de complexité phonologique. Enfin, à partir de la Théorie des éléments (Harris, 1990 ; Backley, 2011), nous explorerons comment les propriétés phonologiques interagissent dans les différentes aphasies et proposerons une explication aux transformations observées.

# 2 La composante phonologique dans l'aphasie

## 2.1 L'aphasie : la nature des contraintes phonologiques

Des études issues des données de l'aphasie portant sur différentes langues – parmi lesquelles l'anglais, le français, l'italien ou encore le néerlandais – permettent de souligner que l'omission (d'un membre consonantique ou bien la réduction d'une séquence consonantique) et la substitution (assimilation, harmonie consonantique) constituent les opérations les plus fréquentes dans l'aphasie lorsqu'il y a un déficit de nature phonologique (Blumstein, 1973, 1978 ; Béland, 1985 ; Valdois et Nespoulous, 1994 ; Nespoulous et al., 1987 ; Béland et Favreau, 1991 ; Den Ouden, 2002 ; Den Ouden et Bastiaanse, 2003, 2005 ; Baqué et al., 2012 ; Prince, 2016).

Blumstein (1973 : 136) souligne qu'elle observe dans ses données une régularité des processus dans l'aphasie où les mêmes phonèmes / syllabes peuvent être affectés dans certaines positions, indépendamment du type d'aphasie. Cette régularité ne relève pas du hasard et provient de la nature même des unités, des contraintes générales qui gouvernent la phonologie des locuteurs : « regardless of the area of brain damage, the more complex phonological structures are impaired and the less complex phonological structures are relatively preserved. » Elle distingue des unités sous les termes « marquées » et « non-marquées » afin de classer celles qui font ou non défaut dans le langage aphasique. Les unités peuvent être « marquées » à différents niveaux. Sur un plan syllabique, les séquences consonantiques sont ainsi plus marquées que les séquences CV, elles sont par conséquent plus souvent altérées (car plus complexes du fait de leur constitution interne, de leur position au sein de la syllabe, et aussi de leur indice de fréquence typologique). Au sein de ces séquences, certaines unités sont également plus marquées que d'autres selon leur position dans la chaîne. La coda est ainsi plus marquée que la position d'attaque, donc plus sujette à l'omission. Sur un plan segmental, les fricatives sont plus marquées que les occlusives, et les segments voisés plus marqués que les non-voisés. Tout ceci relève de contraintes phonologiques structurelles et positionnelles. Ainsi, un locuteur anglophone ou francophone souffrant d'une aphasie élidera généralement le segment le plus marqué au sein d'un groupe consonantique. En position initiale ou médiane, une séquence occlusive+liquide comme /dʁ/ souffrira de la perte de sa liquide alors la consonne fricative /s/ sera plutôt élidée dans une séquence s+occlusive (Béland, 1985 ; Valdois, 1987 ; Nespoulous et Moreau, 1997, Prince, 2016). Den Ouden (2002), Den Ouden et Bastiaanse (2003) obtiennent des résultats similaires en néerlandais chez des aphasiques non-fluents et fluents. Les travaux de Béland (1985), ainsi que ceux initiés par Nespoulous et al. (1987) montrent qu'il est fréquent de trouver des locuteurs aphasiques qui réalisent des segments non-marqués à la place des segments marqués – ceci étant surtout le cas dans l'aphasie de Broca. Ces derniers substituent les fricatives par des occlusives, et les segments voisés sont remplacés par leurs homologues non-voisés, comme par ex. dans *brosse*, /**bʁɔs**/ › [**pʁɔs**][2]. Selon Baqué et al. (2012), Marczyk et Baqué (2015) ce phénomène est le résultat de troubles phonétiques, lié à l'atteinte des aspects moteurs, plus ou moins prononcés chez ce type de patient. Ce phénomène semble être

[2] Notons cependant que peu de travaux renseignent à propos des paraphasies affectant les lieux d'articulation, la majorité se consacrant à l'étude des modes d'articulation et du voisement.

corrélé à une systématicité des erreurs et a pour conséquence une certaine régularité dans les types de transformations. Les travaux de Den Ouden (2002) et de Den Ouden et Bastiaanse (2003) renvoient à des résultats similaires en néerlandais, où, d'après eux, les contraintes de marque sont davantage actives après la lésion chez les aphasiques de Broca. C'est la raison pour laquelle ils privilégient les structures les moins marquées. Les auteurs insistent cependant sur le fait que l'on rencontre aussi ces phénomènes dans l'aphasie de Wernicke et de conduction.

# 3 Méthodologie et données

## 3.1 Participants

Le corpus étudié comprend les productions orales de vingt locuteurs aphasiques (onze femmes et neuf hommes droitiers, $M_{age}$= 63,10 ans). Tous sont monolingues du français standard. Les locuteurs aphasiques ont été évalué et testé de J+2 à J+25 jours après le traumatisme (AVC). Nous les avons rencontrés et enregistrés en phase subaiguë ou aiguë au sein de l'Unité Neuro-Vasculaire de l'hôpital Nord Laënnec (Nantes, France). Les tableaux aphasiques ont été diagnostiqués à partir de la batterie Montréal-Toulouse 86 et de certains tests complémentaires de la version française du *Boston Diagnostic Aphasia Examination,* HDAE-F réalisée par Mazaux et Orgogozo, 1982[3] L'échantillon retenu pour cette étude est composé des tableaux cliniques suivants : sept locuteurs aphasiques de Broca (dont quatre souffrant d'un déficit moteur, d'une apraxie), six sont aphasiques de Wernicke, quatre aphasiques de conduction, trois souffrent d'une aphasie transcorticale sensorielle. Tous souffrent d'une lésion hémisphérique gauche, de troubles de la production orale et en particulier, d'un déficit phonético-phonologique. Nous observons les données des sept locuteurs aphasiques de Broca et des six locuteurs aphasiques de Wernicke.

## 3.2 Procédure et matériels

L'étude est composée de quarante items comportant une séquence consonantique RT ou sT en position initiale, médiane et finale[4]. Les items sont réalisés lors d'une tâche de dénomination d'image (p.ex. *ca**rt**able*, *e**sc**a**rg**ot, ca**sq**uette, **sc**arabée*). D'autres items retenus ont été réalisés lors de conversation spontanée. Les données ont été transcrites et codées sous le logiciel PHON édité par Rose, MacWhinney, Burn et al. (2006) puis analysées.

## 3.3 Résultats

De nombreuses transformations ont été réalisées. Elles correspondent à des omissions (53,38% du total des transformations), des substitutions (28,94%), des métathèses (3,86%), des réductions totales (13,18%) ou encore des épenthèses (0,64%). Nous ne présentons ici que les substitutions réalisées par les locuteurs souffrant d'une aphasie de Broca et de Wernicke. Ces dernières sont essentiellement consonantiques. Conformément aux travaux de Blumstein (1973, 1978) et Nespoulous et al. (1987), nous avons classé les types de substitutions selon qu'elles affectaient (i) le lieu d'articulation, (ii) le mode, (iii) le voisement, ou encore (iv) le lieu et le mode ou le voisement, c'est-à-dire des transformations qui correspondent à une modification qui concerne plusieurs valeurs pour un seul et même segment.

---

[4] En français, un item ne peut débuter par une séquence consonantique de type /ʁ+occlusive/. Ces séquences se retrouvent uniquement en position médiane et finale de mot.

### 3.3.1. Les substitutions

À partir des données récoltées en conversation spontanée ou lors d'une tâche de dénomination nous avons relevé trente-huit substitutions chez les locuteurs aphasiques de Broca[5]. Cependant, seules huit erreurs ont été réalisées par des locuteurs ne souffrant pas d'une apraxie de la parole, ces derniers ayant réalisés peu de transformations. Vingt-six substitutions sont observées chez les aphasiques de Wernicke. Une observation phonologique fine des réalisations permet de montrer plusieurs faits intéressants.

Sur la totalité des substitutions, toutes ou presque correspondent le plus fréquemment à une modification du lieu d'articulation. 77,70% des substitutions impliquent un changement de lieu chez les aphasiques de Broca, 2,8% impliquent un changement de mode, 16,7% induisent une modification de lieu et/ou de mode en plus du voisement et 2,8% correspondent à des cas de dévoisement. Chez les locuteurs aphasiques de Wernicke, 35,72% correspond aux substitutions de lieu, 17,86% à des substitutions de mode, et 46,42% des substitutions impliquent une modification de plusieurs valeurs. Ces dernières correspondent à un changement qui impute la structure consonantique. L'ensemble des locuteurs aphasiques – indépendamment de la nature de l'aphasie – ne réalisent que très peu de transformations de type vocalique et celles-là ne concernent que les voyelles orales /a/, /i/ et /o/ qui deviennent schwa ou /a/. Autrement dit, si jamais une voyelle est modifiée, elle tend à être substituée par une voyelle plus neutre.

Contrairement aux résultats des études de Baqué et al. (2012) et Marczyk et Baqué (2015) la valeur de voisement ne pose pas d'importante difficulté ici, elle est le plus souvent préservée que d'autres modalités. Seulement trois cas de dévoisement sont recensés : *barbe*, /baʁb/ › [baχ**p**], et, deux exemples qui ne concernent pas une modification du groupe consonantique : *dent*, /**d**ɑ̃/ › [**t**ɑ̃], *garage*, /**g**aʁaʒ/ › [**t**aʁaʒ], tous concernent cependant les aphasiques de Broca[6].

Les transformations de lieu d'articulation sont, au contraire, très fréquentes. Il est frappant de constater qu'elles conduisent le plus souvent à des antériorisations ou à des assimilations avec antériorisation. Par exemple, tous les locuteurs aphasiques de Broca remplacent les occlusives vélaires par des coronales comme l'illustrent les exemples ci-dessous :

| Items | Cibles | Réalisations | Items | Cibles | Réalisations |
|---|---|---|---|---|---|
| 1. *casquette* | /**k**as**k**ɛt/ | › [**t**as**t**ɛt] | 2. *crayon* | /**k**ʁejɔ̃/ | › [**t**χɛlɔ̃] |
| 3. *arc-en-ciel* | /aχ**k**ɑ̃sjɛl/ | › [aχ**t**ɑ̃sjɛl] | 4. *marque-page* | /maχ**k**əpaʒ/ | › [maχ**t**əpaʒ] |
| 5. *escargot* | /eskaʁ**g**o/ | › [**d**esaʁ**d**o] | 6. *guitare* | /**g**itaʁ/ | › [**d**itaʁ] |
| 7. *ordinateur* | /ɔʁ**d**inatœχ/ | › [ɔʁ**n**inatœχ] | 8. *écharpe* | /e**ʃ**aχp/ | › [e**s**aχp] |

On remarque dans les exemples ci-dessus (1,2,3,4) que /k/ est toujours remplacé par [t], chez tous les locuteurs, et en (5 et 6) /g/ par [d]. Un patient substitue systématiquement tous les phonèmes /k/ par des /t/ et ne réalise plus aucune dorsale, notons qu'il souffre également d'une apraxie. On retrouve aussi chez Prince (2011) un exemple très concret de ce cas où les items *crocodile*, *crapaud* et *casquette* sont réalisés de la manière suivante : *crocodile*, /**k**χokodil/ › [**t**χo**t**odil], *crapaud*, /**k**χapo/ › [**t**χapo] ou encore *casquette* est réalisé /**k**as**k**ɛt/ › [**t**as**t**ɛt]. Ces exemples portent sur les séquences TR, c'est-à-dire les attaques branchantes. On remarque des substitutions en position initiale au sein d'un groupe consonantique mais aussi à l'isolée en position initiale,

[3] L'étude présentée ici se voudra essentiellement qualitative, c'est à travers l'analyse fine des opérations que nous observons certains indices sur l'origine du trouble. Nous renvoyons à Prince (2016) pour les résultats quantitatifs.

[6] Toutefois, conformément à Marczyk et Baqué, 2015 ; Baqué et al. 2012, nous soulignons qu'une analyse perceptive demeure insuffisante et que des analyses fines du signal acoustique doivent être maintenant réalisées afin d'étayer ces résultats.

médiane ou finale, et ce, quelle que soit la complexité de la position syllabique. Pour ces cas, c'est donc la nature du segment dorsal qui pose problème. Ce type de substitution, très spécifique, qui concerne le passage d'un lieu dorsal vers un lieu coronal, est propre à l'aphasie de Broca et aux patients souffrant d'une apraxie dans notre étude. Dans de plus rares cas, les labiales deviennent elles aussi des coronales : *parking,* [**p**aχkiŋ] › [**t**aχtin], [paχ**t**iŋ], *aspirateur,* [as**p**iʁa**t**œʁ] › [as**t**iʁatœʁ], [as**t**iʁa**s**œʁ], *inter alia*. Dans le dernier exemple, le mode change également, mais la valeur coronale est toujours privilégiée. Par exemple, la fricative, /s/, dans l'item *serpent,* /**s**ɛχpɑ̃/ › [**t**ɛχpɑ̃], se transforme en une occlusive [t]. Autrement dit, nous remarquons que l'occlusivisation d'une fricative porte toujours sur une coronale : *serviette,* /**s**ɛʁvjɛt/ › [**t**ɛʁvjɛt]. Dans *moustique*, /**m**usti**k**/ › [**n**usti**t**], on observe que la nasale bilabiale peut être réalisée comme une dentale nasale. La nasalité perdure alors que le lieu d'articulation change lui aussi pour devenir coronal. Ces faits sont marquants chez les aphasiques de Broca. La systématicité des transformations semble indiquer la présence d'un trouble phonétique et articulatoire, c'est-à-dire que la planification motrice pour les dorsales semble faire défaut. C'est la raison pour laquelle ces dernières sont antériorisées. Or, dans de plus rares cas, il arrive également qu'une coronale devienne aussi une vélaire, tel que dans *cartable*, /kaʁ**t**ablə/ › [kaʁ**k**ab] qui subit une assimilation de lieu par harmonie consonantique, *ordinateur,* /ɔʁ**d**inatœʁ/ › [ɔʁ**k**inatœʁ] et *stylo,* /s**t**ilo/ [s**k**ilo]. Ces faits se présentent plus rarement, seuls quatre exemples sont relevés chez deux aphasiques de Broca. Si ces patients produisent /k/ malgré leur aphasie de Broca et l'apraxie, l'on peut douter d'une atteinte strictement phonétique pour ces cas.
L'un d'entre eux réalise également deux exemples de fricativisation avec voisement avec le mot *serpillière,* /sɛχ**p**ijɛʁ/ › [sɛʁ**v**ijɛʁ], préservant ainsi le lieu labial. Il est le seul patient aphasique de Broca à avoir réalisé cela et l'opération semble ici conduire à une structure plus complexe.

Si la systématicité est moindre et les résultats plus hétérogènes chez les locuteurs souffrant d'une aphasie de Wernicke, des résultats similaires sont cependant observés. On constate de plus nombreuses substitutions de mode mais aussi des substitutions de lieu. Celles-ci vont dans la même direction, l'antériorisation est le plus souvent le résultat des transformations. Les dorsales deviennent coronales et si les coronales subissent un changement un mode, elles conservent leur statut de coronale : *casquette*, /kas**k**ɛt/ › [skas**t**ɛt], *scarabée*, /s**k**aʁabe/ › [s**t**aʁabe], *ordinateur*, /ɔʁ**d**inatœχ/ › [**z**ɔʁ**n**o**z**atœχ], ou encore *cartable*, /kaʁ**t**ablə/ › [kaʁ**n**a**b**lə], [kaʁ**n**a**t**]. Le même patient peut aussi produire des dorsales et donc des phonèmes postérieurs dans le même mot, à l'exemple du premier /k/ de *casquette* qui est correctement réalisé, ou encore de la production de l'item *escargot* réalisé [e**ʃ**kaʁ**g**o] sans modification de la dorsale [g] mais avec la substitution de la fricative alvéodentale non-voisée /s/ en une fricative non-voisée palatale [ʃ]. Quelques exemples illustrent encore l'hétérogénéité des transformations lors de l'aphasie de Wernicke : /t/ est modifié à plusieurs reprises, il devient successivement [ʃ] dans l'item *artichaut*, /aʁ**t**iʃo/ réalisé [aχ**ʃ**iʃo], qui résulte d'une assimilation avec harmonie consonantique ; /p/ dans [saχ**p**eʃo] ; ou encore [l] dans *carton*, /**k**aχ**t**ɔ̃/ › [**ʃ**aʁ**l**õ]. Un exemple a aussi été rencontré avec une consonne labiale dans : *sport*, /s**p**ɔʁ/ › [s**k**ɔʁ], ce résultat peut toutefois s'attribuer à une paraphasie de type lexical.
L'observation fine d'un petit nombre de cas est pertinente et intéressante lorsque l'on considère la dichotomie phonétique-phonologique et l'origine des déficits dans l'aphasie. On remarque ici qu'il n'est pas possible de décrire ces phénomènes si l'on considère qu'il existe une frontière entre les dimensions phonétique et phonologique. Nous avons vu que le contraste entre les lieux coronal et dorsal, par exemple, est bien souvent problématique chez l'ensemble de ces patients, ce que peu d'études relatent à ce jour. Cette tendance est extrêmement forte chez les locuteurs souffrant d'une aphasie de Broca, mais elle ne constitue pas pour autant l'indice d'un trouble d'origine essentiellement phonétique car elle n'est pas systématique et qu'elle apparaît également chez les locuteurs aphasiques de Wernicke. En ce sens, il faut envisager qu'il y a une relation étroite et constante entre plan phonétique et phonologique. C'est la raison pour laquelle on observe bien

souvent des transformations semblables, celles-ci répondant à une logique qui peut s'expliquer à la lumière d'une théorie qui prend en compte cette interface.

### 3.3.2. Le rôle de la phonologie : théorie des éléments

Les éléments sont les unités fondamentales des constituants phonologiques, ils correspondent aux plus petites unités qui caractérisent les segments. Chaque élément correspond à une valeur, en termes d'articulation de lieu, de mode et de voisement ; la somme des éléments propres à la définition d'un segment constitue une matrice. La matrice d'un segment renvoie à l'ensemble des propriétés qui le définissent. À partir de la combinaison d'éléments de base, appelées primitives, il est ainsi possible de décrire tous les segments des langues. Dans le système de représentation de Backley (2011), par exemple, les éléments sont des unités abstraites fondées sur les informations phonologiques utilisées entre locuteurs et auditeurs et directement liées à certaines propriétés acoustiques du signal. Les éléments sont des unités privatives et monovalentes, elles permettent de décrire les systèmes dans les langues et donnent une mesure de la complexité, par exemple : plus un segment comporte d'éléments, plus il est marqué. Chez Backley (2011), six éléments basiques permettent de décrire tout le système. Les propriétés acoustiques et phonologiques pour chaque élément sont définies par : les éléments de mode : |ʔ| = occlusives orales, glottales, |H| = haute fréquence, aspiration, |L| = voisement, nasalité ; et les éléments de lieu : |I| = coronalité, palatalité |A| = pharyngalité, uvularité, et |U| = labialité et vélarité. Dans ce cadre, un segment tel que /t/ reçoit la matrice suivante : |ʔI|, [occlusif, non-voisé, coronal] /d/ recevra la même matrice plus l'élément propre au voisement, |ʔLI|, etc.

Afin de rendre compte du comportement structurel des différents lieux d'articulation et de leur complexité dans l'aphasie, nous proposons ici une version de la théorie des éléments fondées sur la privativité qui répond aux données présentées. Dans cette version, les locuteurs aphasiques, pour qui les segments marqués sont problématiques, remplacent les unités marquées par leurs équivalences non-marquées, autrement dit, par les segments contenant le moins d'éléments. Lors de la sélection des unités phonologiques, l'aphasie entraîne une sélection des unités non-voisées en priorité. Sur un plan phonétique, la caractéristique de voisement n'est plus réalisée, le sujet revient à une position articulatoire neutre, la non-vibration des cordes vocales. L'élément de voisement est donc absent.

Contrairement au voisement, la substitution de lieu correspond à un changement d'élément et non pas à une perte chez Backley. Le passage de /k/ vers /t/ se traduit par une sélection des mauvais éléments, un passage de |U| à |I| qui conduit à une antériorisation lors de la réalisation articulatoire du segment. Nous proposons d'adapter la théorie au français et aux données de l'aphasie. Ici, les coronales sont constituées de l'élément |I|, les labiales de l'élément |U| et les vélaires de l'union des éléments de palatalité |I| et de labialité |U| (conformément à Tifrit, 2013 ; Prince, 2016). Ces dernières sont plus complexes car elles comportent un plus grand nombre d'éléments de lieu. Ainsi, pour décrire les changements entre /p/, /t/ et /k/, on pourrait interpréter le changement de /k/, occlusive vélaire, vers [t], occlusive coronale, comme le passage d'une matrice vélaire |UIʔ| à un élément coronal |Iʔ| *via* la perte de l'élément de labialité |U|, présent dans /k/. Cette perte entraîne une modification lors de sa réalisation articulatoire. Les locuteurs aphasiques de Broca et de Wernicke privilégient les coronales qui sont moins marquées que les dorsales, c'est-à-dire qui contiennent un nombre inférieur d'éléments. Toutefois, et à ce stade, il faut continuer à explorer cette voie où il est nécessaire de rendre compte de la position des labiales par rapport à celles des coronales, les deux comportant le même nombre d'éléments dans le cadre de Backley (2011) mais ne reflétant pas un comportement totalement similaire dans les observations menées ici.

Nous retiendrons que le déficit observé s'envisage comme la perte d'éléments qui conduit à un changement de l'unité lors de la sélection des phonèmes. Cette sélection favorise les unités les moins marquées et conduit, sur le plan moteur, à une réalisation articulatoire déficiente.

# 4 Discussion

Qu'il soit question d'une difficulté de coordination des mouvements phonatoires sur le plan de la planification motrice ou bien d'un accès erroné aux représentations sous-jacentes qui conduit à la sélection de mauvaises unités, les locuteurs souffrant d'une aphasie – de Broca ou de Wernicke – appliquent, de manière inconsciente, des mécanismes de réparation pour parvenir à accéder et réaliser des structures – segmentales et/ou syllabiques. Ces mécanismes sont explicables seulement si l'on accepte la nature d'une relation étroite entre plans phonétique et phonologique.

Indépendamment des tableaux cliniques observés ici, les substitutions réalisées permettent de tirer plusieurs conclusions. Si le type de substitution, très spécifique, qui concerne le passage d'un lieu dorsal vers un lieu coronal est propre à l'aphasie de Broca et aux patients souffrant d'une apraxie, il n'est pas uniquement le résultat d'un trouble moteur et articulatoire. La comparaison avec des patients atteints de l'aphasie de Wernicke permet de montrer qu'ils réalisent eux aussi des transformations similaires, même si ces dernières sont plus hétéroclites. En outre, des aphasiques de Broca ne souffrant pas d'une apraxie réalisent également les mêmes types de transformations.

La sélection des unités phonologiques est connexe à la planification de la production de celles-ci. Si l'étude fine de ces cas à la lumière d'une théorie telle que la théorie des éléments permet de conclure à de premières hypothèses et montre qu'il n'y a pas de réelle frontière entre le plan phonétique et phonologique, de nouvelles recherches, basées sur des analyses phonétiques doivent être maintenant conduites. Conformément aux travaux de Marczyk et Baqué (2015), Baqué et al. (2012), nous suggérons qu'une analyse du signal acoustique des productions des locuteurs souffrant d'une aphasie de Broca et de Wernicke doit être maintenant réalisées afin d'approfondir ces résultats. Nous nous attendons à ce que les aphasiques de Broca souffrant d'un déficit sur le plan moteur réalisent également des transformations imputées à la structure phonologique et aux contraintes pesant sur celle-ci (interaction entre le type de segment, la nature des séquences consonantiques[7], la position de celles-ci au sein de la chaîne syllabique).


# Remerciements

Merci à D. Benichou et aux patients du CHU Laënnec de Nantes pour leur participation à cette étude, ainsi qu'à A. Tifrit, J. Harris, P. Backley & K. Nasukawa et aux relecteurs pour leurs précieux commentaires.


[7] Notons que d'importantes différences ont été relevées entre les substitutions réalisées au sein de séquences /ʁ+occlusive/ et de séquences /s+occlusive/. Autrement dit, d'autres facteurs rendent aussi compte des tendances observées et ne sont pas étudiés ici faute d'espace (distribution des séquences consonantiques, nature hétérosyllabique *vs* tautosyllabique, fréquence des segments, *inter alia*). Ces différences sont liées à des contraintes phonologiques et phonétiques qui sont discutées dans Prince 2016.

# Références


ALAJOUANINE T, OMBREDANE A ., DURAND M. (1939). *Le syndrome de désintégration phonétique dans l'aphasie*. Paris : Masson.

BACKLEY P. (2011). An introduction to Element theory. Edinburgh: Edinburgh University Press.

BAQUÉ L., MARCZYK A., ROSAS A., ESTRADA M., LE BESNERAIS M. NESPOULOUS JL. (2012). De la matière phonique à la structuration phonologique dans l'aphasie. Calvo M.V. Murillo J. (eds.), *Perception phonique et parole*. Mons : CIPA. 75-98.

BÉLAND R., FAVREAU Y. (1991). On the special status of Coronals in Aphasia. Paradis C., Prunet J.F. (eds.) *Phonetics and Phonology vol.2. The special status of coronals: internal and external evidence*. New York: Academic Press. 201-221.

BÉLAND R. (1985). Contraintes syllabiques sur les erreurs phonologiques. Thèse de doctorat. Université de Montréal. Canada.

BLUMSTEIN S. (1978). Segment structure and the syllable in aphasia. Bell A., Hooper J-B. (eds.), *Syllables and segments*. Holland: North-Holland. 189-200.

BLUMSTEIN S. (1973). A phonological investigation of aphasic speech. The Hague: Mouton.

BUCKINGHAM H., BUCKINGHAM S. (2015). Phonological Disorders. International Encyclopedia of the Social and Behavioral Sciences. 2nde edition. Mons : Elsevier. 45-65.

BUCKINGHAM H.W., CHRISTMAN S. (2008). Disorders of phonetics and phonology. Stemmer B. Whitaker H.A (eds.), Handbook of the Neuroscience of Language. London: Academic Press Elsevier. 127-136.

DEN OUDEN DB. (2011). Phonological Disorders, in *Continuum Companion to Phonology,* in Botma B., Kula N-C., Nasukawa K. (eds.) New-York: Continuum. 320-240.

DEN OUDEN DB., BASTIAANSE R. (2003). Syllable structure at different levels in the speech production *process: Evidence from Aphasia*. Van de Weijer J., van Heuven V.J., Van der Hulst H. (eds.), *Current Issues in Linguistic Theory, The Phonological Spectrum. Vol. II: Suprasegmental Structure*. John Benjamins. 234. 81-107.

DEN OUDEN DB. (2002). *Phonology* in *Aphasia, syllables and segments in level-specific deficits*. Ph.D. Dissertation, *Grodil 39*: University of Groningen.

HARRIS J. (1990). Segmental complexity and phonological government. *Phonology 7* Cambridge: Cambridge University Press. 255-300.

NESPOULOUS J. (2006). Le langage et les processus cérébraux : apport de la linguistique et de la psycholinguistique à l'aphasiologie et à la neuropsycholinguistique cognitive du XXème siècle. Auroux S., Koerner E.F.K. Niederehe H-J., Versteegh K. (eds.) Berlin-New York : History of Language Sciences : Walter de Gruyter. 2671- 2682.

NESPOULOUS JL., JOANETTE Y, SKA B., CAPLAN D., ROCH-LECOURS A. (1987). Production deficits in Broca's and Conduction aphasia: repetition vs reading. *Motor and sensory processes of language*. Keller E., Gopnik M. (Eds.) Hillsdale, N-J : Lawrence Erlbaum Associate Inc. 53-81.

MARCZYK A. (2015). Déficits de la composante phonético-phonologique dans l'aphasie et stratégies compensatoires. Analyse acoustique et perceptive de productions consonantiques de sujets hispanophones. Thèse de doctorat. Université de Barcelone.

MARCZYK A., BAQUÉ L. (2015). Predicting segmental substitution errors in aphasic patients with phonological and phonetic encoding impairments. Loquens. 2. 2. 1-15.

PRINCE T. (2016). Représentations segmentales et syllabiques dans l'acquisition du langage et dans l'aphasie, les séquences sT du français. Thèse de doctorat, université de Nantes.

TIFRIT A. (2013). Contraste, cases vides Le cas des obstruantes du français. Phonologie, Morphologie, Syntaxe, mélanges offerts à Jean-pierre Angoujard. Presses Universitaires de Rennes : Rennes. 157-177.

VALDOIS S., NESPOULOUS JL. ([1994]1999). Partie 2 : Altérations spécifiques des composantes du langage. 1. Perturbations du traitement phonétique et phonologique du langage. Séron X., Jeannerod M. (eds.). *Neuropsychologie Humaine*. Bruxelles : Mardaga. 360-374.

# Description automatique du taux d'expression des femmes dans les flux télévisuels français

David Doukhan Jean Carrive
Institut national de l'audiovisuel (Ina), 4 avenue de l'Europe, 94366 Bry-sur-Marne cedex, France
ddoukhan@ina.fr, jcarrive@ina.fr.fr

RÉSUMÉ

Une approche automatique, fondée sur l'estimation du temps de parole par genre, est proposée pour décrire le taux d'expression féminine dans les flux audiovisuels. Des réseaux de neurones convolutionnels (CNN) sont utilisés pour segmenter les flux audio en zones de musique et de parole, attribuées à des hommes ou des femmes. Le taux d'expression féminine est décrit sur plus de 170 000 heures de flux, correspondant à 21 chaînes de télévision nationale, analysées de 2010 à 2017, ainsi que sur 24 chaînes de télévision régionales, analysées en 2016. Cette description montre que le temps de parole est majoritairement attribué à des hommes dans la télévision française, mais que le pourcentage de parole attribué aux femmes a significativement progressé en huit ans pour au moins sept chaînes.

ABSTRACT

**Automatic description of female speaking time percentage in French TV streams**

This paper presents an automatic approach, based on speaker gender detection, aimed at describing female speaking time percentage in audiovisual streams. Convolutional Neural Network models (CNN) are used to segment audio streams into music and speech segments, attributed to male and female speakers. Female speaking time percentage is described using 170.000 hours of raw streams corresponding to 21 national TV channels, analysed between 2010 and 2017, and 24 regional TV channels, analyzed in 2016. This description shows that speech time is mainly attributed to male speakers in French TV, and that female speaking time percentage has statistically increased in 8 years for 7 channels.

MOTS-CLÉS : Détection du genre du locuteur, Humanités Numériques, Égalité des sexes, ConvNet.

KEYWORDS: Speaker Gender Detection, Digital Humanities, Gender Equality, ConvNet.

# 1 Introduction

La mesure de la place accordée aux femmes dans les médias a été traitée à l'aide d'un grand nombre de méthodologies. A l'échelle européenne, elle a été décrite à l'aide du pourcentage de femmes accédant aux plus hauts postes décisionnels dans les médias par type de métier ainsi que du pourcentage de sujets féminins traités dans les nouvelles (Europe, 2015). Elle a également fait l'objet d'études réalisées à l'échelle mondiale (Macharia, 2015), de rapports publics (Reiser & Gresy, 2008), ou encore d'analyses réalisées par le Conseil supérieur de l'audiovisuel (CSA), en charge de veiller à une plus juste représentation des femmes et des hommes dans les programmes audiovisuels depuis 2014 (CSA, 2017). La place des femmes dans les médias y a été décrite à l'aide du *taux de présence* (pourcentage de femmes présentes à l'antenne), du *taux d'expression* (pourcentage de temps de parole


10.21437/JEP.2018-57

attribué aux femmes), ainsi que du *taux d'identification* (pourcentage de mentions orales du nom de personnalités féminines). Ces taux peuvent être exprimés de manière globale, ou analysés séparément en fonction des parts d'audience ou du statut des locuteurs : invité politique, journaliste...

La mesure manuelle du *taux d'expression* des femmes dans les médias est coûteuse et les études s'y consacrant ont jusqu'ici été réalisées sur des corpus de taille limitée. Le corpus GMM, constitué à partir d'émissions diffusées le 15 mai 2008 sur 6 chaînes de télévision et 6 stations de radio, est composé d'extraits d'une durée variant entre 6 minutes et 3 heures par chaîne (Reiser & Gresy, 2008). L'étude menée par le CSA belge rapporte des temps de parole observés sur un total de 36 heures d'archives collectées sur une période d'une semaine sur 26 chaînes de télévision (Levant *et al.*, 2014). Ces limitations provoquent des biais d'analyse correspondant aux particularités du contexte politique et social dans lequel les programmes ont été collectés. Ces études, basées sur l'analyse d'une période restreinte, sont systématiquement accompagnées d'une description détaillée des événements relatifs à cette période : élections, mouvements sociaux, manifestations sportives et culturelles... Ces éléments étant nécessaire pour caractériser le biais affectant les descripteurs de la place accordée aux hommes et aux femmes dans les médias.

L'approche proposée dans cet article consiste à décrire l'évolution du taux d'expression féminine à l'aide de systèmes de segmentation automatique, permettant d'analyser une masse de données beaucoup plus importante, et ainsi de réduire les effets de biais liés aux contexte de collecte. Ils permettent également d'étendre les analyses à l'ensemble des programmes diffusés et ainsi de mettre en évidence un certain nombre de phénomènes susceptibles d'orienter des analyses qualitatives.

La détection du genre d'un locuteur est rendue possible à l'aide d'un certain nombre de caractéristiques acoustiques. La parole des femmes est généralement caractérisée par une fréquence fondamentale plus élevée, des formants localisés dans de plus hautes plages de fréquences, ainsi qu'une qualité de voix plus *soufflée* (breathy). Les différences prosodiques existant entre la parole des hommes et des femmes varient en fonction des pays : cela signifie que les marqueurs acoustiques du genre dans la parole ne sont pas uniquement physiologiques, mais également appris dans un contexte socio-culturel donné (Pépiot, 2015). La détection du genre est plus difficile pour les locuteurs ayant un fort accent (étranger ou régional), une voix associée à des plages de fréquence fondamentale extrêmes, ou encore s'exprimant à l'aide de motifs prosodiques non standard (voix très expressive, imitation, etc...).

Les systèmes de détection du genre du locuteur sont utilisés depuis de nombreuses années dans les moteurs de transcription automatique de la parole (Lamel & Gauvain, 1995). La détection du genre n'y est pas forcément considérée comme une fin en soi, mais plutôt comme un moyen d'améliorer la qualité de la transcription en sélectionnant le modèle acoustique le plus approprié. Plus récemment, des systèmes de détection du genre ont été utilisés pour améliorer les tâches de reconnaissance d'émotions (Xia *et al.*, 2014), ainsi que les stratégies d'interaction d'interface homme-machine (El Shafey *et al.*, 2014). La détection du genre du locuteur présente encore un certain nombre de difficultés, notamment lorsqu'il s'agit de personnes âgées ou d'enfants (Schuller *et al.*, 2013).

# 2 Système de Segmentation audio

Le système proposé pour la segmentation des flux audio est composés de 3 modules principaux : un détecteur d'activité (seuil énergétique), un système de segmentation parole/musique et un système de segmentation homme/femme. Il est disponible sur le répertoire GitHub de l'Ina sous license open-

source [1]. Les systèmes de segmentation sont modélisés à l'aide de réseaux de neurones convolutionnels composés de 5 couches convolutionnelles et 4 couches denses. Ils utilisent le Mel-Frequency Cepstrum (MFC) du signal audio extrait avec SIDEKIT (Larcher *et al.*, 2016), en utilisant un banc de 24 filtres espacés entre 100 et 8000 Hz, ainsi qu'une taille de pas et de fenêtre de 10 et 25 millisecondes. L'entrée des réseaux de neurones est composée de fenêtres de 68 MFC (contexte de 680 millisecondes, 68*24 dimensions) normalisées localement en soustrayant leur moyenne et en les divisant par leur écart type. La sortie des systèmes correspond à une probabilité d'émission par classe (homme-femme ou parole-musique) qui est estimée toutes les 20 millisecondes, et est fournie en entrée d'un modèle de Markov caché à deux états (HMM), chargé de transformer ces prédictions brutes en segments.

Les modèles de détection du genre ont été entraînés à l'aide du dictionnaire de locuteurs interne de l'Ina (Salmon & Vallet, 2014), qui est à notre notre connaissance la plus grande base de locuteurs annotée manuellement à partir de données audiovisuelles (TV et radio). Ce corpus contient environ 32 000 extraits de parole, attribués à 1 780 hommes (94 heures) et 494 femmes (27h). Ces extraits ont été obtenus à partir d'archives collectées entre 1957 et 2012. L'exhaustivité de cet ensemble d'apprentissage fournit un avantage technologique certain pour la conception de systèmes automatiques, et permet d'entraîner des modèles de reconnaissance plus complexes (nombre de paramètres libres, profondeur) en limitant les risques de sur-apprentissage.

La description détaillée et l'évaluation du système ont été réalisées lors de précédents travaux, où ses performances avaient été comparées à des systèmes fondés sur l'utilisation de *i-vecteurs* ainsi que de modèles de mélanges de gaussiennes (Doukhan *et al.*, 2018a). Le système CNN était associé à de meilleurs résultats bruts (F-Mesure de 96.52%), ainsi qu'à une meilleure capacité à estimer le taux d'expression dans des extraits de durée variant entre 30 et 60 minutes (erreur moyenne : 0.59%, erreur maximale : 1.8%). Il a également été montré que plus la durée de l'enregistrement considéré est longue, plus l'erreur d'estimation du taux d'expression est basse. Suite à cette étude, le système a été jugé suffisamment robuste pour pouvoir décrire de manière fiable les variations du taux d'expression.

# 3 Variations du taux d'expression dans les flux TV nationaux

## 3.1 Corpus d'étude

La table 1 présente les 21 chaînes sélectionnées pour décrire les variations du taux d'expression dans les flux télévisuels nationaux. Cette sélection est composée de 7 chaînes publiques et 14 chaînes privées. Elle a été réalisée de manière à faire apparaître les chaînes associées aux plus fortes audiences, ainsi que certaines chaînes thématiques cibles (informations, sport, histoire, musique, contenu visant un public féminin).
Les analyses ont porté sur l'ensemble des flux diffusés de 2010 à 2017. Afin de réduire le temps de calcul, les flux ont été découpés en tranches d'une heure, et un échantillon de 20% des segments a été sélectionné aléatoirement pour être analysé. La plage horaire d'analyse a été restreinte de 10h du matin à minuit, ce qui correspond aux créneaux pour lesquels les audiences de la télévision sont supérieur à 10% (CSA, 2015). La quantité d'échantillons analysés correspond au nombre de chaînes (21) multiplié par le nombre d'heures analysées par jour (14h), le nombre de jours (8 ans) ainsi que le taux d'échantillonnage (20%) ; soit environ 170 000 heures de contenu (20 ans de flux brut).

1. `http://github.com/ina-foss/inaSpeechSegmenter`

| Code | Nom | Statut | Contenu |
|---|---|---|---|
| ART | Arte | Public | Chaîne franco-allemande à vocation culturelle |
| BFT | BFM TV | Privé | Information nationales en continu |
| C+ | Canal+ | Privé | Généraliste axée sur le cinéma et le sport |
| C25 | Chérie 25 | Privé | Généraliste visant un public féminin |
| E21 | L'Équipe TV | Privé | Thématique à contenu sportif |
| ESP | Eurosport | Privé | Thématique à contenu sportif |
| F24 | France 24 | Public | Chaîne publique d'information internationales en continu diffusée en 4 langues et dans 180 pays |
| FR2 | France 2 | Public | Généraliste. Deuxième chaîne la plus regardée en France |
| FR3 | France 3 | Public | Généraliste à vocation régionale : 24 éditions régionales, 44 éditions locales, et 6 éditions en langues régionales |
| FR5 | France 5 | Public | Généraliste axé sur l'éducatif et les documentaires |
| FRO | France Ô | Public | Généraliste consacré à la France d'outre-mer |
| HIS | Histoire | Privé | Thématique consacré à l'histoire |
| ITL | I-Télé | Privé | Information nationale en continu |
| LCI | La Chaîne Info | Privé | Information nationale en continu |
| LCP | LCP/Public Sénat | Public | Politique (Assemblée nationale et Sénat), information |
| M6 | M6 | Privé | Généraliste. Troisième chaîne la plus regardée en France. |
| N12 | NRJ 12 | Privé | Généraliste axée sur les divertissements |
| TEV | Téva | Privé | Généraliste visant un public féminin et familial |
| TF1 | TF1 | Privé | Généraliste. Chaîne la plus regardée en Europe |
| TMC | TMC | Privé | Généraliste |
| W9 | W9 | Privé | Généraliste à vocation musicale et de divertissement |

TABLE 1 – Chaînes de télévision sélectionnées pour décrire décrire le taux d'expression par genre

## 3.2 Description Globale

La figure 1 présente une visualisation globale des 21 chaînes du corpus selon deux modalités : le pourcentage de parole (100 - pourcentage de musique), et le taux d'expression féminine (100 - pourcentage de parole masculine), observés entre 2010 et 2017.
Le pourcentage de paroles varie entre 61 et 93,7%. Il est minimal pour W9 (chaîne musicale) et maximal pour l'ensemble des chaînes d'information, ainsi que les chaînes de sport.
Le taux d'expression des femmes varie entre 7,73 et 47,44%, ce qui signifie que dans la totalité des chaînes analysées les hommes ont un temps de parole supérieur aux femmes. Il est minimal dans les chaînes ayant une programmation sportive (Eurosport, L'Équipe, et dans une moindre mesure CANAL+), et légèrement inférieur à la moyenne dans les chaînes à programmation culturelle (Histoire, Arte, France 5). Les chaînes d'information en continu privées (I-Télé, LCI, BFM-TV) ont des propriétés similaires (taux de parole compris entre 89,9 et 90,6%, pourcentage de parole féminine compris entre 33,4 et 36,4%). Quatre chaînes sont associées à un taux d'expression des femmes supérieur à 40% : les deux chaînes visant un public féminin (Téva et Chérie 25), France 24 et M6.

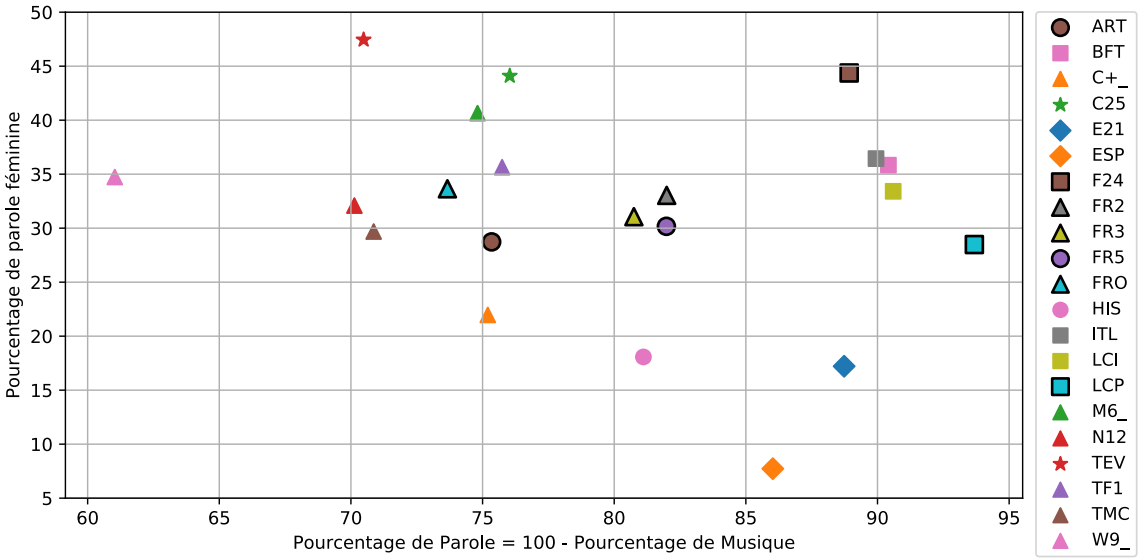



FIGURE 1 – Pourcentages de parole et taux d'expression des femmes observés à la télévision de 2010 à 2017. Les marqueurs à bordure noire correspondent aux chaînes publiques, les triangles aux chaînes généralistes, les cercles aux chaînes à contenu culturel ou éducatif, les losanges au contenu sportif, les carrés aux chaînes d'information et politiques, les étoiles aux chaînes visant un public féminin.

| chaîne | FR5 | W9 | M6 | ITL | TF1 | FR2 | LCP | TEV |
|---|---|---|---|---|---|---|---|---|
| pente | 1.29 | 0.89 | 0.52 | -1.16 | 0.67 | 1.05 | 0.68 | 1.29 |
| p-score | 3.9e-04 | 7.6e-04 | 2.2e-03 | 3.3e-03 | 7.3e-03 | 8.3e-03 | 8.7e-03 | 3.3e-02 |

TABLE 2 – Évolution du pourcentage de parole des femmes entre 2010 et 2017 pour les chaînes ayant une pente statistiquement significative ($p\text{-}score < 0.05$)

## 3.3 Évolution du taux d'expression des femmes de 2010 à 2017

La figure 2 présente l'évolution du pourcentage de parole féminine pour 9 chaînes du corpus associées à des fortes audiences. Pour l'ensemble de ces chaînes, le taux de parole féminine estimé en 2017 est supérieur à celui estimé en 2010. Les chaînes associées aux plus fortes évolutions sont France 2 qui passe de 27,7 à 37,2% et France 5 qui passe de 27,4 à 35,1%.

La table 2 présente les tendances associées cette évolution. Une procédure de régression linéaire (`scipy.stats.linregress`) a été utilisée pour associer à chaque chaîne une mesure de la pente du pourcentage de parole féminin, ainsi qu'un p-score obtenu à l'aide du test de Wald. Huit chaînes sur 21 ont été associées à une évolution significative du pourcentage de parole féminine ($p\text{-}score < 0.05$). La chaîne d'information I-Télé est la seule qui est associée à une diminution du temps de parole des femmes (-1.16% par an). Les trois chaînes associées aux plus fortes croissantes du taux d'expression des femmes sont France 5, Téva, et France 2.

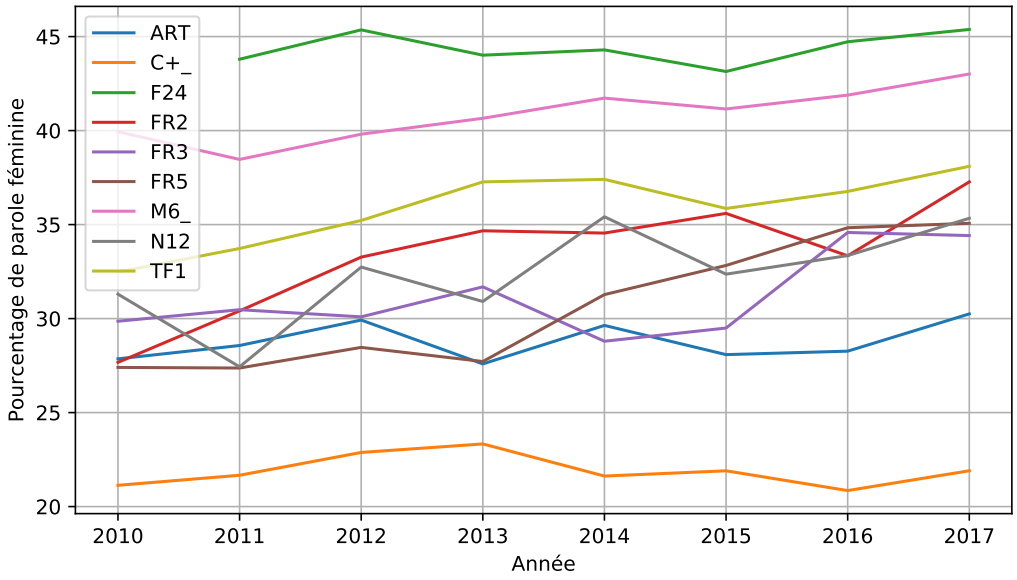



FIGURE 2 – Évolution du taux d'expression des femmes de 2010 et 2017

# 4 Variations du taux d'expression dans les flux TV régionaux

## 4.1 Corpus d'étude

L'étude des disparités régionales du taux d'expression par sexe a été réalisée en analysant la totalité des éditions régionales du 19/20 de France 3 diffusées en 2016. Le 19/20 est une émission d'information diffusée en *prime-time* sur France 3, associée à des fortes parts d'audience, pouvant varier de 14 à 21% (Ozap, 2010; Guadalupe, 2016; Meffre, 2017).

Les 24 éditions régionales du 19/20 sont diffusées simultanément sur France 3, de 19h à 19h30. Elles peuvent être entrecoupées de programmes de publicité ou de bulletins météo, et sont suivies par une édition nationale du 19/20. Les éditions régionales correspondent au découpage en 21 régions métropolitaines de la France datant d'avant 2016, auxquelles s'ajoute la Corse. La région Provence-Alpes-Côte d'Azur a deux éditions distinctes (Provence-Alpes, Côte d'Azur), ainsi que la région Rhône-Alpes (Rhône, Alpes).

## 4.2 Sélection des données

Les éditions régionales ont été isolées dans les flux à l'aide d'une méthode automatique de traitement d'image relativement simple (CNN), consistant à identifier la bannière associée à l'ensemble des journaux régionaux (figure 3). Cette stratégie permet de détecter avec précision l'heure de début et de fin de l'édition, et d'exclure de l'analyse les programmes non désirés : publicité, météo, édition spéciale diffusée sur l'ensemble des chaînes, édition nationale, émission de substitution. Chaque édition a été associée à 5 jours et demi de données brutes, soit 4 mois de flux en considérant l'ensemble des régions, qui ont été analysés dans leur intégralité.


FIGURE 3 – Capture d'écran d'une édition régionale du 19/20 sur France 3, contenant quasi-systématiquement une bannière intitulée 19/20 suivi du nom de la région ou localité considérée

## 4.3 Disparités régionales

La figure 4 détaille le pourcentage de parole attribué aux femmes dans les 24 édition régionale du 19/20. Ce pourcentage varie entre 25.89% et 52.9%. L'Alsace et et Nord-Pas-de Calais sont les seules éditions pour lesquelles le temps de parole attribué aux femmes est supérieur au temps attribué aux hommes. Sept éditions sur 24 sont associées à un temps de parole par sexe à peu près égal (compris entre 45 et 55%), à savoir Alsace (52,90%), Nord-Pas-de-Calais (50,67%), Ile-de-France (47,25%), Picardie (47,13%), Bretagne (46,34%), Provence-Alpes (46,31%), Languedoc-Roussillon (46,06%). Quatre éditions régionales sont associées à un temps de parole des femmes inférieur à un tiers : Lorraine (25,89%), Midi-Pyrénées (31,25%), Auvergne (32,72%), Aquitaine (32,91%).

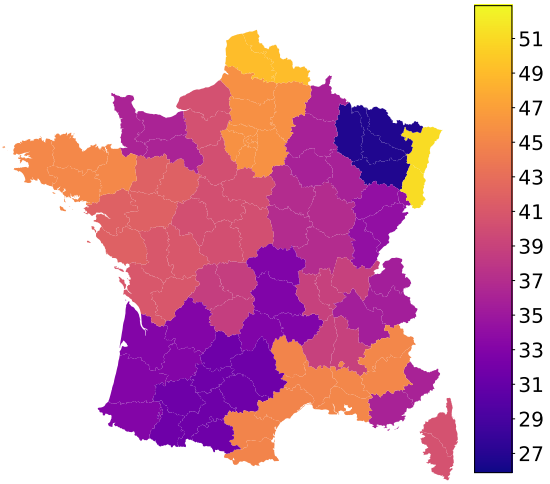



FIGURE 4 – Pourcentage de temps de parole attribué aux femmes dans les éditions régionales d'actualités du 19/20 de France 3 en 2016

Une étude de corrélation préliminaire a été réalisée en croisant le pourcentage de parole des femmes avec le nombre d'habitants par département. L'estimation de la corrélation a été réalisée en utilisant le test non-paramétrique de Spearman (Zwillinger & Kokoska, 2000). Une corrélation positive modérée ($\rho = 0.453$) et statistiquement significative ($pvalue < 10^{-5}$) tend à montrer que plus un département est peuplé, plus le taux de parole des femmes est élevé dans le 19/20.

# 5 Biais d'analyse

Les biais liés à cette approche automatique sont liés à la fiabilité du système de détection du genre du locuteur, qui n'a été entraîné et évalué que sur des voix d'adultes. Ces biais ont été réduits en excluant de l'analyse les chaînes thématiques ciblant les enfants, ainsi que les tranches horaires fréquemment associées à la diffusion de dessins animés (6-9h du matin). La prise en charge des voix d'enfants n'est pas aisée, et il existe assez peu de ressources audiovisuelles permettant d'entraîner et évaluer de tels systèmes (Schuller *et al.*, 2013). De plus, les voix d'enfants diffusées dans les dessins animés, dans les programmes doublés en français, ou dans la publicité radiophonique, sont très majoritairement interprétées par des acteurs professionnels adultes, n'ayant pas nécessairement le même sexe que le personnage qu'ils incarnent. Des observations informelles tendent à montrer que les *« voix d'enfant »* sont tantôt considérées comme de la musique (personnages de dessins animés très théâtraux), ou comme des voix de femme.

# 6 Conclusion

Cette étude présente un système d'analyse automatique de la parole fondé sur l'utilisation de réseaux de neurones convolutionnels capables de segmenter des archives audiovisuelles en zones de musique et en zones de paroles attribuées à des hommes ou a des femmes. Ce système a permis de décrire les différences de traitement existant dans les médias entre les hommes et les femmes, via l'estimation du *taux d'expression*. Il a été montré que sur la totalité des chaînes nationales, le temps de parole est attribué dans une plus grande proportion à des hommes, plus particulièrement dans les chaînes à contenu sportif ou culturel. Cette observation corrobore d'autres études concluant que *Le sérieux d'Arte se fait donc avec les hommes ; l'émotion de M6 se fait avec les femmes* (Reiser & Gresy, 2008). Une augmentation statistiquement significative du taux d'expression féminine a été observé sur 7 chaînes nationales entre 2010 et 2017, ce qui laisse supposer qu'une partie ces inégalités de traitement tend à se réduire.

Cette étude constitue à notre connaissance le premier cas d'usage de systèmes de reconnaissance automatique du genre du locuteur appliqué à la description du taux d'expression féminine. Il s'agit également de la description réalisée sur la plus grande quantité de données, correspondant à plus de 170 000 heures de flux télévisuels. Cette masse de données permet de réduire le biais associé au contexte de collecte des archives mentionné dans les analyses réalisées via des processus d'annotation manuelle (Reiser & Gresy, 2008; Levant *et al.*, 2014).

Les travaux en cours consistent à approfondir les analyses présentées, en y incluant des mesures d'audience par tranche horaire, ainsi qu'une sélection de stations radio (Doukhan *et al.*, 2018b). Des travaux de plus long terme consisteraient à décrire les différences de traitement des hommes et des femmes dans les médias à l'aide d'un plus grand nombre de descripteurs. L'utilisation de systèmes de transcription automatique de la parole devrait permettre d'estimer assez facilement les *taux d'identification*, et pourrait dans une moindre mesure aider à déterminer le statut des locuteurs (expert, témoin, etc...). L'estimation des *taux de présence* pourrait également être envisagée en se basant sur des procédures de segmentation et regroupement en locuteurs.

# Références


CSA (2015). Les chiffres clés de l'audiovisuel français, édition du premier semestre.

CSA (2017). La représentation des femmes à la télévision et à la radio - rapport sur l'exercice 2016.

DOUKHAN D., CARRIVE J., VALLET F., LARCHER A. & MEIGNIER S. (2018a). An open-source speaker gender detection framework for monitoring gender equality. *Acoustics, Speech and Signal Processing (ICASSP)*.

DOUKHAN D., POELS G. & CARRIVE J. (2018b). Describing gender equality in french audiovisual streams with a deep learning approach (accepted). *Journal of European Television History and Culture (VIEW)*.

EL SHAFEY L., KHOURY E. & MARCEL S. (2014). Audio-visual gender recognition in uncontrolled environment using variability modeling techniques. In *International Joint Conference on Biometrics (IJCB)*, p. 1–8 : IEEE.

EUROPE (2015). Toolkit » sur la mise en application de la recommandation du comité des ministres du conseil de l'europe cm/rec (2013) 1 sur l'égalité entre les femmes et les hommes et les médias. *Conseil de l'Europe, Commission pour l'égalité entre les femmes et les hommes*.

GUADALUPE F. (2016). Audiences access : Le "19/20" de france 3 leader, "c à vous" et "28 minutes" en forme. `http://www.ozap.com`.

LAMEL L. F. & GAUVAIN J.-L. (1995). A phone-based approach to non-linguistic speech feature identification. *Computer Speech & Language*, **9**(1), 87–103.

LARCHER A., LEE K. A. & MEIGNIER S. (2016). An extensible speaker identification sidekit in python. In *Acoustics, Speech and Signal Processing (ICASSP)*, p. 5095–5099 : IEEE.

LEVANT B., HANNOT M. & DERINOZ S. (2014). How gender representations matter with generation in television ? In *II International Conference Gender and Communication*, p. 17–27.

MACHARIA S. E. A. (2015). *Who Makes the News ? : Global Media Monitoring Project 2015*. World Association for Christian Communication.

MEFFRE B. (2017). Audiences access : Nagui leader en baisse, "le 19/20" devant "dna", "c à vous" en forme. `http://www.ozap.com`.

OZAP (2010). France 3 : Le 19/20 au dessus des 20% de parts d'audience cette semaine. `http://www.ozap.com`.

PÉPIOT E. (2015). Voice, speech and gender :. male-female acoustic differences and cross-language variation in english and french speakers. *Corela. Cognition, représentation, langage*, (HS-16).

REISER M. & GRESY B. (2008). L'image des femmes dans les médias. *Secrétariat d'Etat à la solidarité*.

SALMON F. & VALLET F. (2014). An effortless way to create large-scale datasets for famous speakers. In *LREC*, p. 348–352.

SCHULLER B., STEIDL S., BATLINER A., BURKHARDT F., DEVILLERS L., MÜLLER C. & NARAYANAN S. (2013). Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, **27**(1), 4–39.

XIA R., DENG J., SCHULLER B. & LIU Y. (2014). Modeling gender information for emotion recognition using denoising autoencoder. In *Acoustics, Speech and Signal Processing (ICASSP)*, p. 990–994 : IEEE.

ZWILLINGER D. & KOKOSKA S. (2000). *CRC standard probability and statistics tables and formulae*. Chapman & Hall : New York.

# Effets de la durée vocalique et du locuteur sur le degré de coarticulation C-à-V en français : étude sur grands corpus.

Fanny Guitard-Ivent[1]
(1) Lab. de Phonétique et Phonologie, UMR 7018, CNRS/Sorbonne-Nouvelle,
19 rue des Bernardins, 75005 Paris, France
fanny.ivent@univ-paris3.fr

## RESUME

Cette étude examine la coarticulation C-à-V en fonction de la durée des voyelles et de la variabilité entre locuteurs à partir de grands corpus du français. Les changements de F2 des voyelles selon le contexte consonantique (alvéolaire *vs.* uvulaire) sont testés sur 26k voyelles /i, E, a, ɔ, u/ (/E/ = /e, ɛ/) réparties en 3 tranches de durée (courte/moyenne/longue). L'effet du locuteur et l'interaction locuteur*durée sont étudiés à partir de 3,5k /a/ produits par 8 locuteurs. Les résultats montrent que plus les voyelles sont courtes, plus l'effet contextuel est fort (excepté pour /u/). La coarticulation comme sa modulation par la durée sont fonction du locuteur. Pour deux locuteurs, l'effet contextuel est indépendant de la durée alors qu'un ajustement sur deux ou trois niveaux de durée est observé pour les autres. Plus surprenant, l'augmentation de la coarticulation des voyelles courtes se fait sur un contexte particulier dépendant du locuteur.

## ABSTRACT

**Effects of vowel duration and speaker on the degree of C-to-V coarticulation in French: study on large corpora.**

This study examines C-to-V coarticulation according to vowel duration and inter-speaker variability using large French corpora. Contextual effects are measured as F2 changes in relation to the adjacent consonant (alveolar *vs.* uvular). 26k tokens of /i, E, a, ɔ, u/ (/E/=/e, ɛ/) splitted into 3 duration classes (short, medium, long) are studied. Speaker-dependent effects and the interaction between speaker and vowel duration, are analyzed from 3.5k /a/ produced by 8 speakers. Results show that the shorter the vowels, the stronger the C-to-V coarticulation (except for /u/). Coarticulation and its modulation according to vowel duration are speaker-dependent. For two speakers, contextual effect is independent of vowel duration whereas two- or three-way contrasts are observed for the others. Surprisingly, the increase of coarticulation for short vowels favors a particular context (alveolar or uvular) which is speaker-dependent.

MOTS-CLES : coarticulation, locuteur, durée, acoustique, grand corpus
KEYWORDS: coarticulation, speaker, duration, acoustic, large corpora

# 1 Introduction

L'étude qui est présentée ici s'insère dans le cadre de mes recherches de doctorat visant à mieux comprendre la coarticulation C-à-V en français en analysant notamment si et comment le contexte consonantique interagit avec d'autres facteurs de variation. Les effets de la coarticulation sur les


10.21437/JEP.2018-58

propriétés spectrales des voyelles sont particulièrement étudiés à partir de l'examen de grands corpus de parole.

Les grands corpus de parole permettent de tester de manière systématique des phénomènes phonétiques, tels que la coarticulation, en prenant en considération différents facteurs de variation. De récentes études, traitant de l'interaction entre coarticulation et prosodie en français, ont montré que la coarticulation C-à-V (Guitard-Ivent, Fougeron, 2017) comme V-à-V (Turco, Audibert, Fougeron, 2015), étudiées à partir de 17k et 33k voyelles respectivement, étaient réduites en position initiale de Groupe Intonatif (comparé à la position interne de Groupe Intonatif). Une autre étude portant sur l'interaction entre coarticulation (C-à-V et V-à-V) et style de parole, menée sur 55k voyelles, a montré que les voyelles étaient plus fortement coarticulées en parole conversationnelle qu'en parole journalistique, plus formelle (Turco, Guitard-Ivent, Fougeron, 2017). Ainsi, les tendances observées en parole naturelle ont permis de confirmer des résultats d'études menées sur du matériel plus contrôlé en condition de laboratoire (p.ex. : Cho et al., 2017 pour l'effet de la position prosodique, et Moon, Lindblom 1994 pour l'effet du style de parole) parfois contredits par d'autres travaux. Ce présent travail est une sous-analyse des études précédemment citées. On sait que la durée des segments et le locuteur sont des facteurs de variation majeurs. L'idée est donc d'aller encore plus au cœur des données afin de tester comment le contexte consonantique interagit avec ces derniers facteurs pour apporter un éclairage nouveau aux principaux résultats de ma thèse, enrichir et affiner certaines interprétations.

La durée des segments est la cause de nombreuses variations dans la parole. On sait notamment que les voyelles courtes sont plus fortement influencées par le contexte consonantique (Lindblom, 1963 ; Vaissière, 1985 ; Gendrot, Adda-Decker, 2005). Selon Lindblom, l'amplitude des déplacements de formants dépend de la durée de la voyelle et de la distance entre le locus des consonnes et la cible des voyelles. Plus la durée est courte et les distances CV sont importantes, plus les déplacements sont importants. Il semble donc crucial de prendre ces deux paramètres en compte pour mieux appréhender les variations vocaliques.

Au delà des variations dues aux caractéristiques anatomiques des locuteurs, chaque individu possède des particularités articulatoires qui lui sont propres. Ainsi, le signal de parole se trouve coloré par les caractéristiques idiosyncratiques du locuteur, comme les spécificités dues à la provenance géographique du locuteur ou à son appartenance sociale, pouvant se révéler être des indices pertinents pour la reconnaissance du locuteur. Plusieurs études affirment que la coarticulation V-à-V (Grosvald, 2009) comme C-à-V (Su et al., 1974) seraient fonction du locuteur, mais les résultats concernant la pertinence de cet indice pour la reconnaissance du locuteur divergent. Su et al. (1974) affirment que la coarticulation est un indice pertinent pour la reconnaissance du locuteur mais cela a été contredit par Khan (2011). Selon Grosvald et Su et al. nous nous attendons à observer des différences entre locuteurs dans le degré de coarticulation C-à-V indépendamment des éventuelles implications que cela peut avoir pour la reconnaissance du locuteur.

On sait que les formants des voyelles courtes n'atteignent pas leur cible acoustique et que la durée est le facteur le plus déterminant pour les cas d'*undershoot* (Lindblom, 1963). Si, nous trouvons des locuteurs qui coarticulent moins que d'autres, on peut se demander s'ils s'accommodent de la durée de la même manière que les autres. Est-ce que des locuteurs qui coarticulent peu, sont aussi moins sensibles aux variations de durée ? Selon Moon et Lindblom (1994) l'amplitude des variations formatiques traduisent d'une certaine manière l'effort articulatoire d'un locuteur. En augmentant la précision articulatoire, et donc l'effort articulatoire fourni, un locuteur peut parler plus vite sans

coarticuler plus. Notre étude porte uniquement sur des données acoustiques, nous ne pourrons donc pas tester directement l'effort articulatoire. Cependant, il sera intéressant de voir si certains locuteurs s'accommodent différemment des variations de durée, et de mettre ces résultats en relation avec les profils coarticulatoire et sociolinguistique du locuteur.

Dans le but de mieux comprendre la coarticulation C-à-V en français et d'affiner les interprétations de nos précédentes études portant sur l'interaction entre coarticulation et prosodie, ou coarticulation et style de parole, nous chercherons à savoir : si et comment la coarticulation C-à-V est modulée par la durée des voyelles ; 2) si certains locuteurs se distinguent dans leur profil coarticulatoire ; et 3) si la modulation de la coarticulation selon la durée de la voyelle est fonction du locuteur.

# 2 La coarticulation C-à-V est-elle fonction de la durée vocalique ?

## 2.1 Méthode

L'ensemble du matériel linguistique est extrait de deux corpus du français standard publiquement disponibles : ESTER (Gravier et al., 2006) et NCCFr (Torreira et al., 2010). ESTER est un corpus de parole journalistique contenant des émissions radio-télévisées, en partie pré-écrites, d'informations et de débats sur des sujets politiques et sociétaux. Le corpus NCCFr est composé de conversations entre amis sur des sujets sociétaux lors d'échanges en face à face. Pour cette étude, 23 locuteurs masculins (15 NCCFr et 8 ESTER) ont été sélectionnés selon la quantité de matériel à disposition.

### *2.1.1 Matériel linguistique & prétraitements*

À partir des données de ces 23 locuteurs, nous avons analysé 26k exemplaires des voyelles /i, E, a, u, ɔ/ (/E/=/e, E/) en séquences CVC. La consonne adjacente (gauche ou droite) pouvait être soit une consonne alvéolaire ($C_{ALV}$=/t, d, z, s, l, n/, p.ex. *dépanner* /depane/) connue pour attirer F2 vers leur locus à 1800Hz), soit une consonne uvulaire ($C_{UV}$=/R/ p.ex. *appareil* /apaRɛj/, connue pour abaisser le F2 des voyelles (et élever F1). Le contexte opposé était toujours une consonne labiale. Enfin, selon d'autres études sur le français menées sur des corpus semblables (Gendrot, Adda-Decker, 2005) ou identiques (Audibert et al., 2014), les voyelles étaient regroupées en trois catégories de durée : 1) les voyelles d'une durée de 50 ms étaient rangées dans la catégorie des voyelles courtes (COURTE) ; 2) les voyelles de 60 à 80 ms étaient assignées la catégorie voyelles moyennes (MOYENNE) et 3) les voyelles dont la durée était comprise entre 90 et 150 ms faisaient partie des voyelles longues (LONGUE). Un résumé du matériel linguistique utilisé est présenté Tableau 1.

Les effets contextuels ont été mesurés par les changements du second formant (F2) de la voyelle V selon le lieu d'articulation de la consonne adjacente C. Les valeurs de F2 ont été extraites en utilisant Praat (Boersma, Weenink, 2014) sur un alignement automatique forcé, en prenant la moyenne des valeurs extraites à 1/3, 1/2 et 2/3 de chaque voyelle. Afin d'éliminer les valeurs de formants aberrantes, un filtre a été selon les mêmes critères que ceux utilisés dans (Gendrot, Adda-Decker, 2005). Afin de réduire les potentiels effets dus aux différences anatomiques entre locuteurs, une transformation des valeurs de F2 en z-scores par locuteur a été effectuée.

### *2.1.2 Analyse statistique*

Afin de tester l'effet de la durée sur la coarticulation C-à-V, nous avons construit un modèle mixte pour chacune des 5 voyelles étudiées /i, E, a, ɔ, u/ à l'aide du logiciel R et de la bibliothèque 'lme4'. Ainsi, nous avons testé les relations entre les valeurs de F2 (z-score) de V et les facteurs suivants : 1) *durée* de la voyelle (COURTE *vs*. MOYENNE *vs*. LONGUE) et 2) *contexte* ($C_{ALV}$ *vs*. $C_{UV}$). La structure fixe contenait aussi une interaction entre les deux facteurs. Nous avons modélisé un intercept par locuteur et par mot. Afin d'éviter un taux élevé d'erreur de Type I les pentes aléatoires par locuteur et mot ont été incluses pour le facteur *contexte*. Cela correspond à la variabilité entre locuteurs (et entre mots) de l'effet du contexte sur F2. Les valeurs de p ont été obtenues par approximations de type *Satterthwaite* à l'aide de la fonction *lmerTest*. Le seuil de référence a été fixé à $p<.05$. Les effets de chaque facteur fixe et de leur interaction ont été testés par comparaison de modèles avec la fonction *anova*. Les valeurs de $R^2$ associées à chaque modèle ont été obtenues à l'aide de la fonction *r.squaredGLMM* intégrée dans la bibliothèque 'MuMIn'. Les analyses des contrastes à postériori ont été effectuées avec la fonction *lsmeans* (bibliothèque 'emmeans').

| | COURTE | | MOYENNE | | LONGUE | |
|---|---|---|---|---|---|---|
| | $C_{ALV}$ | $C_{UV}$ | $C_{ALV}$ | $C_{UV}$ | $C_{ALV}$ | $C_{UV}$ |
| /i/ *(4048)* | 1094 | 140 | 1862 | 262 | 571 | 119 |
| /E/ *(9123)* | 2611 | 744 | 3462 | 1086 | 885 | 335 |
| /a/ *(8616)* | 2389 | 721 | 3341 | 1095 | 735 | 335 |
| /ɔ/ *(2579)* | 471 | 449 | 737 | 533 | 205 | 184 |
| /u/ *(2022)* | 167 | 371 | 378 | 699 | 140 | 267 |
| *total* | *9157* | | *13455* | | *3776* | |

TABLE 1 : Résumé du matériel linguistique.

## 2.2 Résultats

Les résultats des analyses statistiques par voyelle sont résumés dans le Tableau 2. On peut voir qu'un effet du *contexte* consonantique s'observe sur toutes les voyelles /i,E,a,ɔ,u/. Sans surprise, le F2 des voyelles est plus élevé en contexte alvéolaire qu'en contexte uvulaire. Un effet de la *durée* sur le F2 des voyelles est aussi observé. Comme attendu, plus les voyelles /i/ et /E/ sont longues, plus leur F2 est élevé alors que l'allongement de la voyelle /u/ se traduit par un abaissement de son F2. Ceci indique que ces voyelles tendent plus vers leurs cibles lorsqu'elles sont longues. En revanche, un effet partiel est trouvé pour la voyelle /a/ : le F2 d'un /a/ moyen est plus élevé que le F2 d'un /a/ long. Aucun effet de la durée n'apparait pour /ɔ/. Plus important, une interaction entre nos deux facteurs *contexte* et *durée* a été trouvée pour toutes les voyelles excepté /u/. Plus les voyelles /i,E,a,ɔ/ sont courtes, plus l'effet contextuel est fort. L'effet trouvé s'échelonne sur les trois niveaux de durée testés (COURTE > MOYENNE > LONGUE). En revanche, la coarticulation de la voyelle /u/ se fait indépendamment de la durée de la voyelle. Quelle que soit la durée de /u/, l'effet contextuel reste inchangé. Ces résultats sont présentés Figure 1 (à gauche la voyelle /i/, au centre la voyelle /a/, et à droite la voyelle /u/). Il est intéressant de noter que la voyelle /i/ souvent décrite comme peu ou pas variable présente, comme les autres voyelles, des variations contextuelles. Une résistance à la variation est observée uniquement lorsque la voyelle est longue.

| | Effets des facteurs fixes | | | | Interaction | |
|---|---|---|---|---|---|---|
| | contexte | $\chi^2(1)$ | durée | $\chi^2(2)$ | contexte*durée | $\chi^2(2)$ |
| /i/ $R^2_{(m,c)}$=0.05, 0.32 | ALV > UV | $16^{***}$ | C < M < L | $44^{***}$ | Effet du contexte : C > M > L | $35^{***}$ |
| /E/ $R^2_{(m,c)}$=0.15, 0.50 | ALV > UV | $47^{***}$ | C < M < L | $282^{***}$ | Effet du contexte : C > M > L | $51^{***}$ |
| /a/ $R^2_{(m,c)}$=0.48, 0.68 | ALV > UV | $65^{***}$ | L < M | $11^{**}$ | Effet du contexte : C > M > L | $57^{***}$ |
| /ɔ/ $R^2_{(m,c)}$=0.52, 0.80 | ALV > UV | $50^{***}$ | C = M = L | $0.5^{ns}$ | Effet du contexte : C > M > L | $57^{***}$ |
| /u/ $R^2_{(m,c)}$=0.28, 0.53 | ALV > UV | $58^{***}$ | C > M > L | $102^{***}$ | Effet du contexte : C = M > L | $1^{ns}$ |

TABLE 2 : Tests de vraisemblance des modèles mixtes testant les effets du *contexte*, de la *durée* et leur interaction pour les cinq voyelles testées /i,E,a,ɔ,u/. Les valeurs de $\chi^2$ sont reportées comme estimations de la taille d'effet. Spécifications : pour le facteur *contexte* « ALV » = alvéolaire, « UV » = uvulaire ; pour le facteur *durée* « C » = COURTE, « M » = MOYENNE et « L » = LONGUE. Les valeurs de p significatives sont indiquées par l'astérisque $^{*}$ (p<.05 = $^{*}$, p <.001 = $^{**}$, p<.000 = $^{***}$).

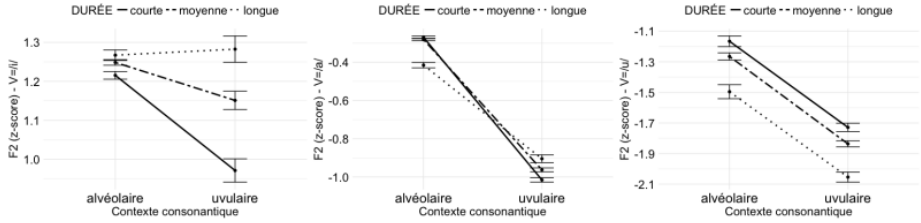



FIGURE 1: F2 (en z-score) des voyelles /i/ (à gauche), /a/ (au centre) et /u/ (à droite) en fonction du *contexte* consonantique (alvéolaire *vs.* uvulaire) et de la *durée* des voyelles (COURTE, MOYENNE, LONGUE).

# 3 Analyse par locuteur

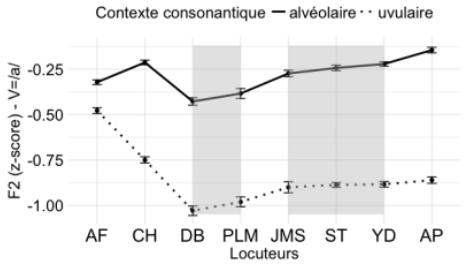
Afin de tester si la coarticulation C-à-V est fonction du locuteur, 3,5k /a/ produits par les 8 locuteurs masculins du corpus ESTER, sélectionnés pour l'étude précédente, ont été analysés. Le choix de la voyelle s'est fait après observation du matériel. La voyelle /a/ étant la plus fréquente dans notre corpus, son analyse rendait l'analyse d'une interaction entre locuteur et durée envisageable.

## 3.1 La coarticulation C-à-V est-elle fonction du locuteur ?

Dans cette partie nous analysons si les changements de F2 dus à la consonne adjacente sont dépendants du locuteur ou non. Autrement dit, nous voulons voir si la coarticulation est fonction du locuteur. Pour cela, nous avons construit un modèle linéaire mixte pour la voyelle /a/ dans le but de tester les relations entre les valeurs de F2 (z-score) et les facteurs *contexte* (alvéolaire *vs.* uvulaire) et *locuteur* (8 niveaux) en interaction. La structure aléatoire est composée d'un intercept par mot et d'une pente aléatoire pour le facteur contexte (ni l'interaction, ni le facteur locuteur n'ont pu être inclus dans la pente aléatoire pour problèmes de convergence). Pour les détails des logiciels et fonctions utilisés voir la sous-section 2.1.2. Les locuteurs sont présentés Figure 2. Nous noterons

que le profil dialectal d'un des huit locuteurs se distingue des autres. En effet, le locuteur AF est un journaliste camerounais. Il sera intéressant de voir si ce locuteur se distingue ou non des autres dans son profil coarticulatoire.

Le modèle mixte ($R^2_m$ = 0.60, $R^2_c$ = 0.79) révèle un effet du contexte ($\chi^2$ (1) = 659, p < .0001$^{***}$) et du locuteur ($\chi^2$ (7) = 330, p < .0001$^{***}$). Le F2 de /a/ est plus élevé en contexte alvéolaire qu'en contexte uvulaire. Le F2 de /a/ diffère d'un locuteur à l'autre sauf dans quelques cas : pas de différence entre les locuteurs CH et PLM, ni entre DB et ST tout comme entre JMS, YD et AP. Plus intéressant, notre analyse fait ressortir une interaction entre nos deux facteurs ($\chi^2$ (1) = 411, p < .0001$^{***}$) montrant que la coarticulation est fonction du locuteur. Les tests d'interactions menés à postériori nous ont permis de répartir les locuteurs sur une échelle de coarticulation sur quatre niveaux : AF < CH < PLM = DB < JMS = ST = YD < AP. La Figure 2 illustre ce résultat. Les locuteurs sont présentés en abscisse et rangés hiérarchiquement du moins « coarticulateur » au plus « coarticulateur ». Les zones grises indiquent qu'il n'y pas de différence significative entre les locuteurs concernés. La ligne pleine représente le contexte alvéolaire et la ligne en pointillés représente le contexte uvulaire. Plus l'écart entre les deux lignes pour un locuteur donné est important, plus l'effet contextuel est fort.




| INITIALES | LOCUTEURS |
|---|---|
| AF | Alain Foka |
| CH | Christophe Hondelatte |
| DB | Dominique Bromberger |
| PLM | Pierre Le Marc |
| JMS | Jean-Marc Sylvestre |
| ST | Simon Tivolle |
| YD | Yves Decaens |
| AP | Alain Passerel |

FIGURE 2: F2 de /a/ en z-score par locuteur (rangé du moins « coarticulateur » au plus « coarticulateur ») en fonction du contexte consonantique (alvéolaire *vs.* uvulaire). Les zones grises indiquent un effet identique du contexte pour les locuteurs concernés. Les noms et prénoms des locuteurs correspondants aux codes locuteurs, sont reportés dans le tableau à droite du graphique.

## 3.2 La modulation de la coarticulation par la durée de V dépend-elle du locuteur ?

L'objectif de cette dernière analyse est de voir si la modulation par la durée est la même pour tous les locuteurs. Pour tester cela, à notre précédent modèle, nous avons rajouté le facteur *durée* en interaction avec le *contexte* consonantique et le *locuteur*. La structure aléatoire reste inchangée à savoir un intercept par mot et d'une pente aléatoire pour le facteur contexte. Nos précédentes analyses ont révélées une interaction : 1) entre le contexte et la durée (pour les voyelles /i,E,a,ɔ/) et 2) entre le contexte et le locuteur (pour la seule voyelle testée /a/). Ainsi nous avons pu voir que la coarticulation était modulée par la durée et qu'elle était fonction du locuteur. Si notre nouvelle analyse, menée sur la voyelle /a/, révèle une triple interaction, nous pourrons conclure que la modulation de la coarticulation selon la durée de la voyelle est fonction du locuteur.

Le modèle mixte ($R^2_m$ = 0.62, $R^2_c$ = 0.79) révèle un effet du contexte ($\chi^2$ (1) = 661, p < 0001$^{***}$), du locuteur ($\chi^2$ (7) = 330, p < .0001$^{***}$) et de la durée ($\chi^2$ (2)=6.607, p=0.04$^{*}$). Comme attendu, le F2 de /a/ est plus élevé en contexte alvéolaire qu'en contexte uvulaire et il diffère d'un locuteur à l'autre. Concernant la durée, l'analyse des contrastes par paire ne révèle au final aucune différence de F2

entre les catégories de durée. Plus intéressant, notre analyse révèle une triple interaction entre nos trois facteurs *contexte*, *locuteur* et *durée* ($\chi^2$ (14)=29.388, p = 0.009**). La modulation de la coarticulation selon la durée de la voyelle est donc fonction du locuteur. L'analyse des contrastes montre plusieurs patterns visualisables Figure 3 : 1) Pour deux locuteurs (CH et YD), la coarticulation est indépendante de la durée de la voyelle ; 2) un locuteur (PLM) ajuste son degré de coarticulation en fonction de la durée sur trois niveaux (COURTE > MOYENNE > LONGUE) ; 3) quant aux cinq autres, une modulation sur deux niveaux est observée. Quatre d'entre eux (DB, JMS, ST et AP) coarticulent plus les voyelles courtes et moyennes que les longues. Alors que le dernier (AF), qui est aussi le locuteur le moins « coarticulateur », coarticule plus les voyelles courtes que les moyennes et les longues. Contre toute attente, l'augmentation de la coarticulation des voyelles courtes se fait sur un contexte particulier dépendant du locuteur. Ce résultat sera discuté dans la prochaine section.

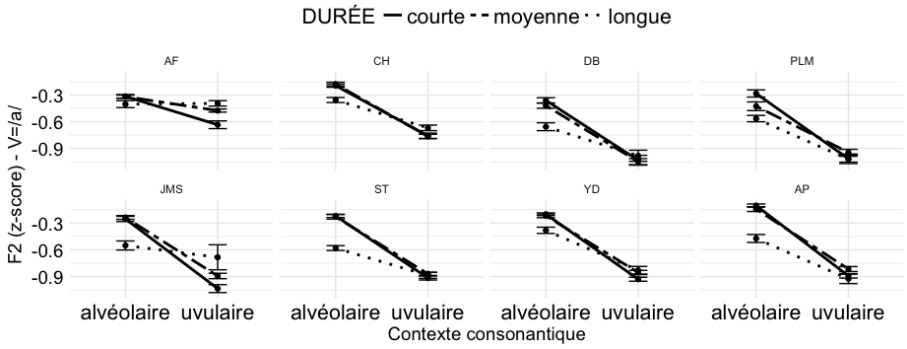



FIGURE 3 : F2 de la voyelle /a/ en z-score en fonction du *contexte* consonantique (alvéolaire *vs.* uvulaire) et de la *durée* de la voyelle (COURTE, MOYENNE, LONGUE) par locuteur.

# 4 Discussion & Conclusion

L'examen de 26k exemplaires des voyelles /i,E,a,ɔ,u/ en séquence CVC a montré que plus les voyelles /i,E,a,ɔ/ étaient courtes plus elles étaient enclines à la coarticulation. L'exploration de grands corpus de parole corrobore les résultats d'études menées sur du matériel plus contrôlé (Lindblom, 1963 ; Vaissière, 1985), et va dans le sens des travaux montrant que les voyelles sont plus variables lorsqu'elles sont courtes (Gendrot, Adda-Decker, 2005 ; Meunier, Espesser, 2012 ; Audibert et al., 2015). Cependant, en ce qui concerne la voyelle /u/, l'effet contextuel reste inchangé quelle que soit la durée de la voyelle. Ce dernier résultat, en opposition avec van den Heuvel et al. (1996), montre que pour cette voyelle les effets contextuels ne se résument pas à des cas de voyelles *undershoot* dus à un manque de temps pour atteindre les cibles articulatoires. Cela est cohérent avec la tendance universelle d'antériorisation des voyelles postérieures fermées pour laquelle l'hypothèse d'une origine coarticulatoire a été émise (Harrington, 2012).

Nos résultats de l'analyse par locuteur sur 3,5k /a/ ont montré que la coarticulation était fonction du locuteur comme cela avait déjà été rapporté (Nolan, 1983 ; van den Heuvel et al., 1996). Nous avons notamment identifié des locuteurs qui coarticulent très peu (AF) et d'autres qui coarticulent beaucoup (ex :AP). Il est intéressant de noter que le locuteur présentant un très faible degré de coarticulation, soit le locuteur camerounais. Ce profil de coarticulation particulier serait-il une spécificité dialectale ? On peut aussi imaginer que ces différences entre locuteurs soient attribuables à des variations de débit. Est-ce que les locuteurs qui coarticulent beaucoup, sont aussi ceux qui parlent vite ? La prise en compte du débit de parole semble nécessaire pour déterminer si ces

différents patterns de coarticulation sont juste la conséquence d'un débit de parole plus ou moins rapide.

La modulation de la coarticulation C-à-V selon la durée de la voyelle s'est révélée être dépendante du locuteur. Sur les huit locuteurs testés, deux ne modulent pas leur coarticulation selon la durée de la voyelle (CH et YD). Cela nous permet d'écarter l'idée selon laquelle un locuteur qui coarticule peu, serait moins affecté par les variations de durée puisque YD coarticule beaucoup (voir Figure 2). De plus, ce résultat semble indiquer que la coarticulation C-à-V n'est pas juste le résultat de voyelles réduites (*undershot)* causées par des contraintes temporelles. Les six autres locuteurs montrent plus de coarticulation sur les voyelles courtes. Alors que notre première analyse montrait que la coarticulation C-à-V était modulée selon les trois niveaux de durée testés (COURTE > MOYENNE > LONGUE), un seul de nos 8 locuteurs (PLM) présente effectivement une gradation du degré de coarticulation sur trois niveaux de durée. La première analyse contenait aussi les données des 15 locuteurs du corpus NCCFr. Est-ce que la modulation de la coarticulation par la durée est plus forte en parole conversationnelle où la coarticulation est plus importante (Turco et al., 2017) ? Une future analyse par locuteur sur ce corpus, permettra de répondre à cette question. Dans tous les cas, pour nos cinq autres locuteurs, la modulation de la coarticulation selon la durée de la voyelle, ne se fait que sur deux niveaux de durée. Pour quatre d'entre eux (DB, JMS, ST et AP) les voyelles courtes et moyennes sont plus coarticulées que les voyelles longues alors que pour un locuteur (AF) seules les voyelles courtes sont plus coarticulées que les voyelles moyennes et longues. La distribution des exemplaires de durée moyenne (60-80ms) nous permet d'écarter l'hypothèse d'une répartition inégale des exemplaires entre nos locuteurs. L'explication de cette différence se trouverait plutôt dans le profil particulier du locuteur AF. Ce dernier, coarticule très peu, les cas d'*undershoot* pour ce locuteur sont moins nombreux. Seules les voyelles les plus courtes sont marquées par un effet contextuel net pour ce locuteur. Dès que ces voyelles dépassent 50 ms, la coarticulation est nettement réduite comme c'est le cas pour les voyelles longues d'autres locuteurs. Cela traduit-il une plus grande précision articulatoire de la part de ce locuteur, suggérant un effort articulatoire plus important comme l'explique Moon, Lindblom (1994) ?

Enfin, il semblerait que l'ajustement coarticulatoire selon la durée de la voyelle se fasse sur un contexte particulier dépendant du locuteur. En effet, excepté pour le locuteur JMS réduisant l'effet de chaque contexte avec l'allongement de la voyelle, pour les cinq autres locuteurs affectés par la durée, l'ajustement se fait sur un seul contexte. Pour quatre d'entre eux, l'allongement de la voyelle conduit à une réduction de l'effet de la consonne alvéolaire. Pour le dernier locuteur (AF), la réduction de la coarticulation sur les cibles plus longues, passe par une élévation du F2 en contexte uvulaire généralement connu pour abaisser F2. L'observation des autres voyelles produites par ce locuteur, montre que l'ajustement de la coarticulation en fonction de la durée de la voyelle, se fait toujours sur ce contexte uvulaire. On peut donc facilement s'imaginer qu'une variation dialectale soit à l'origine de cet ajustement coarticulatoire particulier sur le contexte uvulaire.

Cette étude, menée sur 26k voyelles extraites de grands corpus de parole, a permis d'appuyer des observations menées sur des données plus contrôlées. À savoir que plus les voyelles sont courtes, plus elles sont coarticulées et que la coarticulation est fonction du locuteur. Nous avons aussi montré que la modulation de la coarticulation selon la durée de la voyelle était dépendante du locuteur. Les particularités dialectales permettent d'expliquer les principales variations entre locuteurs. La prise en compte du dialecte dans l'étude de la coarticulation semble donc fondamentale pour la suite.

# Références


AUDIBERT N., FOUGERON C., GENDROT C., ADDA-DECKER M. (2015). Duration- vs. style-dependent variation: a multiparametric investigation. *Proceedings of the 2015 ICPhS* conference.

CHO T., KIM D., KIM S. (2017). Prosodically-conditioned fine-tuning of coarticulatory vowel nasalization in English. *Journal of Phonetics* 64, 71-89.

GRAVIER G., BONASTRE J-F., GEOFFROIS E., GALLIANO S., TAIT K. Mc., CHOUKRI K. (2006). Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news. *Proceedings of European Conference on Speech Communication and Technology*, 139-142.

GROSVALD M. (2009). Interspeaker variation in the extent and perception of long-distance vowel-to-vowel coarticulation. *Journal of Phonetics* 37 (2), 173-188.

GUITARD-IVENT F., FOUGERON C. (2017). Domain-Initial strengthening as reduced coarticulation. *Proceedings of Phonetic and Phonology in Europe 2017 (PaPE)*.

HARRINGTON J. (2012). The coarticulatory basis of diachronic high back vowel fronting, in M., Solé, J., Recasens, D., others (The Initiation of Sound Change: Perception, Production and Social Factors), 103-122.

KAHN J. (2011). Parole de locuteur : performance et confiance en identification biométrique vocale.

LINDBLOM B. (1963). Spectrogaphic study of vowel reduction. *Journal of the Acoustical Society of America* 35, 1773-1781.

MEUNIER C., ESPESSER R. (2012). Vowel reduction in conversational speech in French: The role of lexical factors. *Journal of Phonetics* 39 (3), 271-278.

MOON S. J., LINDBLOM B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America* 96, 40-55.

NOLAN F. (1983). *The Phonetic Bases of Speaker Recognition.* Cambridge : Cambridge Univ. Press.

SU L. S., LI K. P., FU K. S. (1974). Identification of speakers by use of nasal coarticulation. *Journal of the Acoustical Society of America* 56, 1867-1882.

TORREIRA F., ADDA-DECKER M., ERNESTUS M. (2010). The Nijmegen corpus of casual French. *Speech Communication* 52, 201-212.

TURCO G., FOUGERON C., AUDIBERT N. (2016). The effects of prosody on French V-to-V coarticulation : A corpus-based study. *Proceedings of Interspeech*, 998-1001.

TURCO G., GUITARD-IVENT F. FOUGERON C. (2017). Speech style effects on non-local and local coarticulation in French. *Proceedings of International Seminar of Speech Production*.

VAISSIÈRE J. (1985). Étude des variations allophoniques de la voyelle /a/ et ses conséquences pour la reconnaissance automatique de la parole. In XIV Journées d'Etudes de la Parole, Paris.

VAN DEN HEUVEL H., CRANEN B., RIETVELD T. (1996). Speaker variability in the coarticulation of/a, i, u. *Speech communication* 18 (2), 113-130.

# Entre Québec et France, qu'en est-il de l'antériorisation de /ɔ/ en français contemporain ?

Xavier St-Gelais[1], Christophe Coupé[2], François Pellegrino[2] et Vincent Arnaud[1, 2]

(1) Université du Québec à Chicoutimi, 555, boul. de l'Université, Chicoutimi (Québec), G7H 2B1, Canada
(2) Laboratoire Dynamique du Langage UMR5596, Université de Lyon et CNRS, 14, avenue Berthelot, 69007 Lyon, France
xavier.st-gelais1@uqac.ca, vincent.arnaud@uqac.ca

## RESUME

L'antériorisation de /ɔ/, largement étudiée en France, n'a reçu que peu d'attention au Québec. Afin de documenter une éventuelle variation diatopique entre France et Québec, une analyse acoustique comparative de la fréquence centrale du deuxième formant ($F_2$) de 2837 voyelles produites dans des mots et non-mots monosyllabiques (C)VC par des étudiants universitaires de Saguenay (Québec) et de Lyon (France) a été menée. Un modèle de régression linéaire à effets mixtes appliqué aux données indique que /ɔ/ est significativement plus antérieur à Lyon qu'à Saguenay. Dans les deux villes, le lieu d'articulation de la consonne antéposée et celui de la consonne postposée influencent la structure acoustique de cette voyelle. Quelle que soit leur position, les consonnes antérieures (ex. /t, d/) favorisent le $F_2$ le plus élevé ; les consonnes labiales (ex. /p, b/), le $F_2$ le plus bas.

## ABSTRACT

**/ɔ/-fronting in contemporary French between Quebec and France**

/ɔ/-fronting has been widely studied in France, but this phenomenon has received little attention in Quebec French. To better understand a potential regional variation between France and Quebec province, an acoustic analysis of the second formant frequency ($F_2$) of 2837 vowels uttered in (C)VC monosyllabic words and non-words by university students from Saguenay (Quebec) and Lyon (France) was conducted. By fitting a linear mixed effects regression to the data, /ɔ/ is found to be significantly more fronted in Lyon than in Saguenay. In both dialects, the places of articulation of the initial and final consonant also influence the vowel's acoustical structure. No matter their position, front consonants (ex. /t, d/) favour a higher $F_2$, while labials (ex. /p, b/) are associated with the lowest $F_2$.

MOTS-CLES : antériorisation de /ɔ/, français, Québec, sociophonétique, acoustique de la parole, modèle de régression linéaire à effets mixtes

KEYWORDS: /ɔ/-fronting, French, Quebec, sociophonetics, acoustics, linear mixed effects model

# 1 Contexte

En français, la voyelle /ɔ/ est fréquemment soumise à une antériorisation. Remarqué par des grammairiens dès le XVI[e] siècle (voir Fónagy, 1989), ce phénomène a fait l'objet de quantité de


10.21437/JEP.2018-59

travaux depuis l'article fondateur de Martinet (1957). Différents facteurs linguistiques apparaissent influencer l'antériorisation de /ɔ/, comme la graphie (Hansen et Juillard, 2011), la fréquence lexicale (Woehrling et Boula de Mareüil, 2007 ; Mooney, 2016) ou encore l'accentuation : si, dans la littérature, les exemples fournis sont surtout situés en position inaccentuée (Walter, 1976 ; Coveney, 2001), le phénomène peut aussi affecter les voyelles en position accentuée[1] (Lennig, 1979 ; Paradis, 1985 ; Mooney, 2016). Cette dernière position est toutefois plus rarement étudiée.

Des effets des segments adjacents sur l'antériorisation de /ɔ/ sont aussi mentionnés dans la littérature. Woehrling et Boula de Mareüil (2007), dans leur étude acoustique fondée sur deux vastes corpus oraux, indiquent que les consonnes antérieures, qu'elles soient en position antéposée ou postposée, favorisent l'antériorisation de /ɔ/. Les auteurs ne précisent toutefois pas si le contexte syllabique a été contrôlé. Armstrong et Low (2008) soutiennent, quant à eux, que le lieu d'articulation de la consonne postposée dans une syllabe fermée serait le déterminant majeur du degré d'antériorité de /ɔ/ : pour ces auteurs (2008 : 439), en français, « assimilation generally proceeds in an 'anticipatory' or 'regressive' direction ». Par l'entremise d'une analyse auditive visant à évaluer comme antériorisées ou non 372 occurrences produites en parole spontanée, les auteurs notent un taux d'antériorisation de 86,3 % devant des consonnes antérieures (/t, d, s, l, n, ʃ/), mais de seulement 13,3 % devant des postérieures (/k, ʁ/). Les consonnes antérieures impliquant un mouvement de la langue vers l'avant, Armstrong et Low estiment qu'il est logique qu'elles favorisent une antériorisation. Leur analyse souligne aussi que la consonne labiale /m/, qui n'implique pas de mouvement lingual, favorise aussi l'antériorisation. Armstrong et Low (2008 : 441) avancent que la haute fréquence d'apparition de cette consonne après /ɔ/ pourrait expliquer cet effet : « [o]ther things being equal, it appears uncontroversial to state that high frequency will promote the adoption of a linguistic innovation. » Dans son étude acoustique récente en Béarn, chez les locuteurs âgés de 16 à 18 ans, Mooney (2016) relève lui aussi que les consonnes postérieures, en position postposée, sont moins favorables à une augmentation de $F_2$ que les antérieures et les labiales /m, f/. Mooney (2016 : 73) postule que « [t]outes les consonnes qui favorisent une voyelle antérieure, /ʃ, m, t, f, s, l/, comprennent un rétrécissement antérieur, même si les labiales /m, f/ n'ont aucun geste lingual », ce qui expliquerait leur effet similaire. Toujours selon cet auteur, en position antéposée, les consonnes labiales de même que le /ʁ/ sont plutôt associées à des occurrences de /ɔ/ plus postérieures, les consonnes antérieures favorisant quant à elles un $F_2$ plus élevé. L'effet des consonnes labiales serait donc différent avant et après la voyelle.

L'antériorisation de /ɔ/ a également été associée à plusieurs facteurs extralinguistiques. Si Martinet (1957) considère le phénomène comme typique des classes populaires parisiennes, Carton (2000 : 31) y voit une « marque de préciosité inconsciente ». Selon les études et les lieux d'enquête, l'antériorisation de /ɔ/ est associée soit à un âge avancé (Walter, 1976 ; Paradis, 1985), soit à la jeunesse (Armstrong et Low, 2008 ; Lamontagne, 2015 ; Mooney, 2016), au registre spontané (Woehrling et Boula de Mareüil, 2007 ; Woehrling, 2009) ou à la parole lue (Malderez, 1995). Certains auteurs notent aussi que les jeunes femmes pourraient mener le mouvement d'antériorisation de /ɔ/ (Carton, 2000 ; Armstrong et Low, 2008 ; Mooney, 2016).

Au plan géographique, l'antériorisation de /ɔ/ est relevée en Belgique et en Suisse (Woehrling, 2009), de même qu'au Québec (Paradis, 1985 ; Lamontagne, 2015), mais a surtout été étudiée à Paris (Walter, 1976 ; Lennig, 1979) – c'est d'ailleurs dans cette ville qu'elle serait apparue (Martinet, 1957) – et plus généralement en France. Le phénomène serait aujourd'hui en expansion

[1] Étant donné la position de l'accent tonique en français et la *loi de position* régissant en partie la répartition de /ɔ/ et de /o/ en fonction du contexte syllabique, la vaste majorité des /ɔ/ en syllabe accentuée sont situées en syllabe fermée.

dans le français du nord de l'Hexagone (Carton, 2000 ; Woehrling et Boula de Mareüil, 2007). Selon Armstrong et Low (2008), ce trait serait caractéristique d'une prononciation septentrionale et, jouissant d'un certain prestige, se diffuserait par nivellement interdialectal vers d'autres espaces géographiques, comme le Roannais. Mooney (2016) interprète aussi l'antériorisation chez les jeunes Béarnais comme le produit de cette diffusion. Bien que l'antériorisation de /ɔ/ soit mentionnée dans plusieurs points d'enquête, peu d'études ont cherché à évaluer sa variabilité diatopique. En France, Woehrling et Boula de Mareüil (2007) ont comparé les deux premiers formants de voyelles produites par des locuteurs du nord et du sud de la France. Ils ont constaté que l'antériorisation de /ɔ/ était beaucoup moins fréquente en français méridional. Woehrling (2009) a étendu cette analyse en comparant le nord et le sud de la France, l'Alsace, la Belgique et la Suisse : dans ces deux derniers pays, l'antériorisation semble aussi commune qu'au nord de la France, alors qu'en Alsace, elle l'est davantage qu'au Sud, mais moins qu'au Nord. Hors d'Europe, seul Lamontagne (2015) a examiné l'antériorisation de /ɔ/ dans 7 villes du Québec et de l'Ontario. L'auteur note un effet significatif de l'origine géographique sur le phénomène, mais sans plus de précisions. En somme, les données disponibles sont peu nombreuses et peu comparables.

# 2 Objectifs

La présente contribution a trois objectifs : a) fournir des données acoustiques récentes sur la prononciation de la voyelle /ɔ/ en syllabe fermée au Québec et en France, b) vérifier si une variabilité d'origine diatopique s'exprime entre ces deux espaces géographiques, c) confirmer l'effet du lieu d'articulation de la consonne postposée sur la voyelle et explorer celui du lieu d'articulation de la consonne précédente, moins documenté, à la fois au Québec et en France.

# 3 Repères méthodologiques

Les occurrences analysées sont extraites d'un corpus de parole de laboratoire constitué en 2016-2017 auprès de 10 femmes et 9 hommes originaires de la ville de Saguenay, aire urbaine de 160 000 habitants située à 250 km au nord-est de Québec (SG, $\bar{x}$=23,3 ans, $s$=2,6 ans), et de 10 femmes et 9 hommes originaires de Lyon, en France (LY, $\bar{x}$=21,2 ans, $s$=2,3 ans). Tous étaient étudiants universitaires et habitaient dans leur ville d'origine depuis leur naissance au moment de l'enregistrement. La tâche consistait à lire à voix haute, en chambre anéchoïque, des phrases pentasyllabiques terminées par des mots ou des non-mots cibles contenant l'une des voyelles orales du français (ex. « La soupe était *bonne*. »). Les énoncés étaient proposés en ordre aléatoire. Après leur lecture, l'enquêteur feignait de ne pas avoir compris la cible et demandait au locuteur de la répéter isolément. L'utilisation de ce paradigme permettait de désambiguïser les homographes (par ex., <jet>=/ʒɛ/ ou /dʒɛt/) et d'éviter les effets de lecture de listes. Seuls les mots et les non-mots produits isolément ont été considérés.

Parmi les 658 mots et non-mots du corpus destinés à différentes études, 76 (51 mots et 25 non-mots) ayant /ɔ/ pour noyau, tous monosyllabiques et de structure (C)VC, ont été retenus. Le TABLEAU 1 présente leur distribution. Les consonnes initiales et finales ont été étiquetées en fonction de trois lieux d'articulation : labial (/p, m, b, f, v/), antérieur (/t, d, n, l, s, ʃ, ʒ/) et postérieur (/k, g, ʁ/). 2888 occurrences de /ɔ/, toutes accentuées, ont été extraites (76 mots et non-mots × 38 locuteurs). 51 ont été rejetées à cause d'une nasalisation importante (27), d'une erreur de production (20) ou d'une intonation exagérément montante (4). Au total, 2837 voyelles ont été analysées.

| | $C_2$ labiale (/p,b,m,f/) | $C_2$ antérieure (/t,d,n,l,s,ʃ,ʒ/) | $C_2$ postérieure (/g,k/) | Total |
|---|---|---|---|---|
| Aucune $C_1$ | 1 | 3 | 1 | 5 |
| $C_1$ labiale (/p,b,m,f,v/) | 8 | 15 | 5 | 28 |
| $C_1$ antérieure (/t,d,n,l,s/) | 5 | 13 | 5 | 23 |
| $C_1$ postérieure (/g,k,ʁ/) | 5 | 12 | 3 | 20 |
| Total | 19 | 43 | 14 | 76 |

TABLEAU 1 – Tableau de contingence des mots et des non-mots du corpus en fonction du lieu d'articulation des consonnes antéposée ($C_1$) et postposée ($C_2$)

Les voyelles ont été segmentées manuellement à l'aide de PRAAT version 6.0.x[2]. Les frontières ont été placées à des passages par zéro correspondant à l'apparition et à la disparition des patrons vocaliques. Par la suite, à partir de spectrogrammes en bandes larges, les paramètres de détection des fréquences centrales formantiques par LPC (algorithme de Burg) ont été ajustés manuellement pour chaque occurrence ; les trajectoires formantiques obtenues ont été utilisées pour extraire la fréquence centrale du deuxième formant ($F_2$) à 50 % de la durée[3].

Les données ont été analysées avec R version 3.4.x[4], principalement à l'aide de la bibliothèque *lme4* (Bates *et al.*, 2015). Un modèle de régression linéaire à effets mixtes a été construit pour examiner la relation entre la variable dépendante ($F_2$) et les quatre variables indépendantes catégorielles examinées : la ville d'origine (VILLE, deux niveaux), le sexe (SEXE, deux niveaux), le lieu d'articulation de la consonne précédente (PRE_LIEU, quatre niveaux) et celui de la consonne suivante (POST_LIEU, trois niveaux). Toutes les interactions possibles ont également été évaluées. Pour éviter toute pseudo-réplication (Hurlbert, 1984), les facteurs INDIVIDU et MOT ont été intégrés comme effets aléatoires croisés sous forme d'*intercepts* (ordonnées à l'origine). Des pentes aléatoires par INDIVIDU pour les effets de lieu d'articulation (PRE_LIEU et POST_LIEU), corrélées entre elles et avec l'*intercept* pour INDIVIDU, ont aussi été intégrées afin d'éviter une augmentation du risque d'erreur de type I (Barr, 2013).

Afin d'identifier un modèle parcimonieux, les effets fixes ont fait l'objet d'une sélection descendante par estimation de la significativité des différences de déviance entre des modèles successifs nichés (Zuur *et al.*, 2009). Des tests de rapports de vraisemblance (*LR-tests*) ont été utilisés à cette fin, et les variations du critère d'information d'Akaike (*AIC*) ont aussi été observées. La linéarité et l'homogénéité de la variance ont été vérifiées par l'inspection visuelle de diagrammes de dispersion des résidus. Les résidus suivent une distribution normale à queues lourdes symétriques (*symmetrical heavy-tailed*), celle-ci ayant comme seul effet potentiel de rendre le modèle plus conservateur en ce qui a trait à la significativité des effets fixes (Pinheiro et Bates, 2006 : 180). Par ailleurs, les diagrammes quantile-quantile suggèrent que la distribution normale des effets aléatoires ne peut pas être rejetée. Enfin, cinq occurrences estimées trop influentes sur les résultats du modèle de régression ont été exclues. Le pseudo-$R^2$ de Nakagawa et Schielzeth (2013) indique que le modèle final est bien ajusté aux données (pseudo-$R^2_{marginal}$=0,57, pseudo-$R^2_{conditionnel}$=0,85). Étant

[2] BOERSMA P., WEENINK D. (2017). https://www.praat.org/

[3] Dans 306 cas, les trajectoires ne s'ajustaient pas aux formants visibles sur le spectrogramme. Elles ont alors été estimées par interpolation linéaire de points placés manuellement sur le spectrogramme en bande large. Le script conçu est inspiré du travail d'E. Ferragne : https://moodlesupd.script.univ-paris-diderot.fr/mod/page/view.php?id=49768

[4] R CORE TEAM. (2017). https://www.R-project.org/

donné la présence d'interactions (voir section 4), la significativité des effets des prédicteurs a été évaluée par ANOVA de type III reposant sur des tests de Fisher (*F-tests*) et utilisant l'approximation de Kenward-Roger pour estimer les degrés de liberté du dénominateur. Afin d'identifier les niveaux des prédicteurs ayant un impact significatif sur la variable dépendante, des tests *t* de comparaisons multiples avec une correction de Holm fondés sur les moyennes marginales estimées (MME) ont été menés. La significativité des effets a été interprétée en fonction des valeurs exactes de *p*. Le seuil de signification α=0.05 n'a pas été considéré de façon stricte, car comme l'indique Murtaugh (2014 : 613), « [i]t is clear that a decision rule leading to very different interpretations of *P* values of 0.049 and 0.051 is not very rational. » Par ailleurs, les fréquences formantiques brutes n'ont fait l'objet d'aucune normalisation intrinsèque ou extrinsèque. Dans la lignée de Drager et Hay (2012), nous estimons que l'ajustement de l'*intercept* par INDIVIDU, conjointement à l'inclusion du facteur fixe SEXE au sein du modèle, permet de modéliser l'effet du sexe et l'impact de l'individu sur l'ajustement des valeurs formantiques prédites par le modèle.

# 4 Résultats

La FIGURE 1 illustre la dispersion des occurrences de /ɔ/ dans un espace $F_1 \times F_2$ en fonction de VILLE et SEXE. Des différences d'ordre sexuel émergent clairement dans chaque ville. En outre, les voyelles des locuteurs lyonnais semblent plus antérieures que celles des Saguenéens. Le $F_2$ moyen des occurrences produites par les hommes lyonnais est même très proche de celui des /ɔ/ des femmes saguenéennes. Cependant, la forte dispersion des occurrences produites par ce groupe est manifeste.

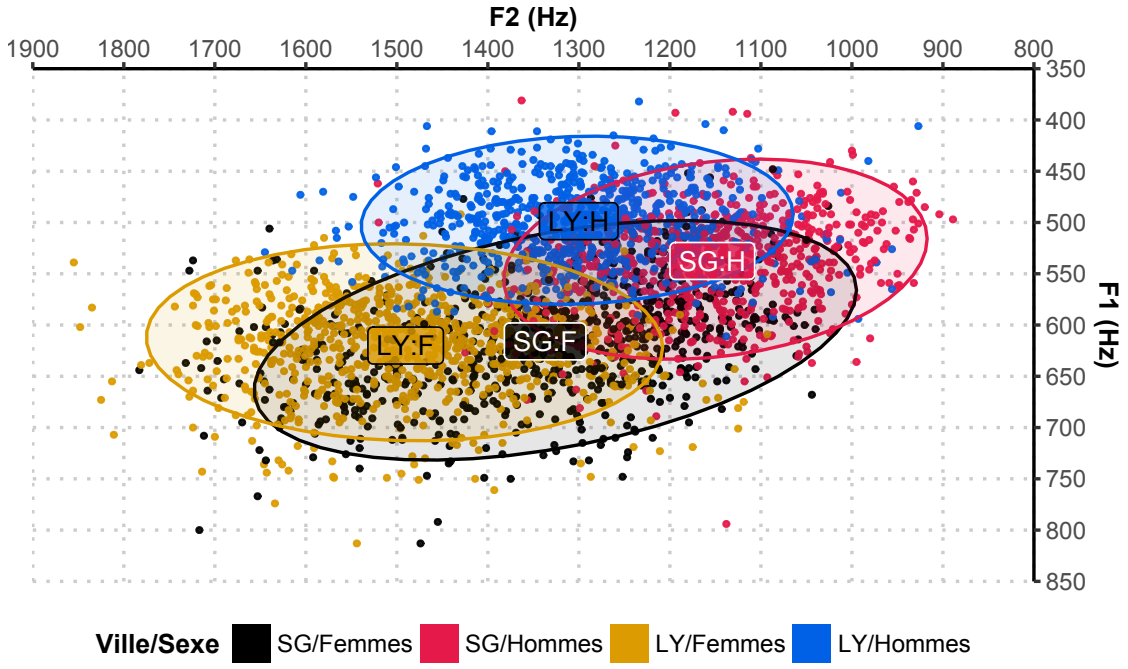



FIGURE 1 – Diagramme $F_1 \times F_2$ (à 50 % de la durée vocalique) des 2837 occurrences de /ɔ/ du corpus (les ellipses de dispersion sont situées à ±2 écarts-types autour des moyennes de groupe)

Au-delà de ces observations préalables, le modèle de régression linéaire à effets mixtes a permis de mettre au jour trois interactions doubles influençant significativement la fréquence de $F_2$ : SEXE × POST_LIEU, VILLE × POST_LIEU et VILLE × PRE_LIEU (FIGURE 2). En premier lieu, l'effet de POST_LIEU sur $F_2$ diffère en fonction de SEXE ($F(2, 33{,}99)=5{,}73$, $p=0{,}007$). Comme attendu, $F_2$ est nettement plus élevé chez les femmes (MME=1385 Hz, IC à 95 % [1350, 1420]) que chez les

hommes (MME=1222 Hz, IC à 95 % [1186, 1259]), et ce, indépendamment du lieu d'articulation de la consonne postposée. Cependant, chez les femmes, les voyelles présentent un $F_2$ significativement plus élevé lorsqu'elles sont suivies de consonnes antérieures que lorsqu'elles sont suivies de consonnes postérieures ou labiales, ces dernières ne présentant pas d'effet différencié. Chez les hommes, une différence significative existe entre consonnes antérieures et consonnes labiales, mais l'effet des consonnes postérieures ne se différencie pas de celui des autres consonnes.

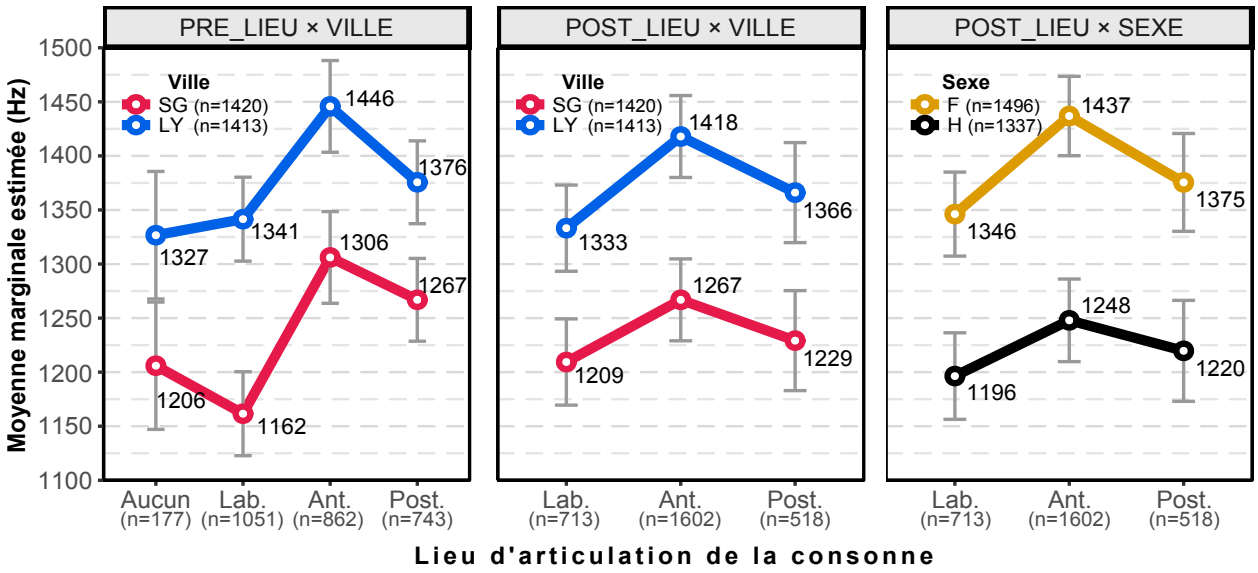



FIGURE 2 – Moyennes marginales estimées de $F_2$ (Hz) et intervalles de confiance à 95 % des trois interactions

Le modèle met également en exergue une tendance de l'effet de POST_LIEU à être modulé par VILLE ($F(2, 34{,}01)=3{,}29$, $p=0{,}049$). À Saguenay comme à Lyon, les consonnes labiales postposées sont associées à un $F_2$ significativement moins élevé que les antérieures ; en outre, dans les deux villes, les effets des consonnes labiales et postérieures ne se distinguent pas. L'interaction est causée par les consonnes postérieures, dont l'effet diffère de celui des antérieures à Lyon ($F_2$ est significativement moins élevé devant des postérieures), mais pas à Saguenay. Par ailleurs, la différence significative liée à la VILLE, sensiblement équivalente d'un contexte consonantique à l'autre, est manifeste.

Cette différence d'ordre géographique est plus marquée en ce qui concerne l'effet différencié de PRE_LIEU en fonction de VILLE ($F(3, 34{,}13)=16{,}08$, $p<0{,}001$). L'ordonnancement des effets du lieu d'articulation de la consonne antéposée est similaire dans les deux villes : $F_2$ est significativement plus élevé après les consonnes antérieures qu'après les labiales ou lorsqu'il n'y a pas de consonne initiale. L'effet de ces deux derniers contextes n'est pas différencié. La différence entre les villes, source de l'interaction, tient aux consonnes postérieures antéposées. À Lyon, $F_2$ dans ce contexte ne diffère pas de celui des voyelles précédées par des consonnes labiales ou sans consonne antéposée, mais est significativement différent de celui des voyelles précédées par des consonnes antérieures. À l'inverse, à Saguenay, l'effet des consonnes postérieures ne se distingue pas de celui des consonnes antérieures, mais se différencie statistiquement de celui des consonnes labiales.

# 5 Discussion et conclusion

L'antériorisation de /ɔ/ apparaît comme plus marquée à Lyon qu'à Saguenay. En moyennant l'effet de SEXE, PRE_LIEU et POST_LIEU, les occurrences des locuteurs lyonnais présentent un $F_2$

(MME=1372 Hz, IC à 95 % [1335, 1409]) plus élevé que celui des voyelles des locuteurs saguenéens (MME=1235 Hz, IC à 95 % [1198, 1272]). La variation diatopique de l'antériorisation de /ɔ/ entre le français québécois de Saguenay et le français hexagonal de Lyon est manifeste.

L'effet du lieu d'articulation des consonnes adjacentes sur le $F_2$ de /ɔ/ est manifeste et, dans l'ensemble, similaire d'une ville à l'autre et d'une position à l'autre. En position antéposée comme postposée, les consonnes antérieures favorisent un $F_2$ plus élevé, tout comme Woehrling et Boula de Mareüil (2007) et Mooney (2016) le mentionnent. Par contre, nos résultats se dissocient de la littérature antérieure concernant les consonnes labiales. Dans la présente étude, elles sont associées au $F_2$ moyen le plus bas dans les deux positions. L'hypothèse articulatoire proposée par Mooney (2016) concernant un « rétrécissement antérieur des consonnes labiales et antérieures » ne peut s'appliquer aux présents résultats. Comme le suggèrent pour leur part Armstrong et Low (2008), le mouvement de la langue impliqué dans l'articulation des consonnes pourrait être responsable du patron observé. Les consonnes labiales, tout comme l'absence de consonne initiale, n'impliquent pas d'intervention de la masse linguale, cette dernière pouvant rester en position postérieure lors de l'articulation vocalique. Si tel est le cas, l'effet des consonnes postérieures à Saguenay demeure difficile à expliquer : elles devraient être associées à un $F_2$ significativement plus bas que les consonnes antérieures, ce qui n'est le cas dans aucune des deux positions. Il est aussi possible que d'autres gestes articulatoires interviennent, notamment l'arrondissement labial. Armstrong et Low (2008) remarquent à cet égard que les occurrences de /ɔ/ perçues plus arrondies ont généralement un $F_2$ plus bas. L'analyse des effets consonantiques sur $F_1$ et $F_3$ pourrait fournir quelques pistes de réponse, mais comme le mentionnent Armstrong et Low, seules des études articulatoires permettraient d'obtenir des conclusions robustes.

Par ailleurs, la présente étude comporte trois limites principales. Premièrement, non-mots et mots n'ont pas été distingués. S'il n'est pas exclu que leur traitement relève de processus cognitifs différents, l'ajout du statut lexical comme facteur fixe dans le modèle statistique indique toutefois que cette variable n'influence pas significativement $F_2$. Deuxièmement, avec une valeur de *p* de 0,049 obtenue par ANOVA de type III, l'interaction POST_LIEU × VILLE apparaît comme une tendance. Afin d'évaluer la robustesse de cet effet, un bootstrap paramétrique reposant sur la construction de 1000 échantillons simulés de données dont la distribution correspond à celle des données originales a été utilisé. Pour chaque échantillon, un *LR-test* visant à tester la significativité de la différence de déviance entre le modèle complet et le modèle n'incluant pas cette interaction a été effectué. La valeur de *p* empirique équivaut à la proportion des valeurs de LR plus grandes ou égales à la valeur de LR observée sur les seules données originales ; elle s'établit à $p$=0,043. L'interaction POST_LIEU × VILLE est donc relativement robuste. Troisièmement, le regroupement des lieux d'articulation des consonnes sous trois étiquettes n'offre qu'un portrait global des effets des contextes consonantiques adjacents et mériterait d'être affiné. Par exemple, Armstrong et Low (2008 : 440) mentionnent qu'un /m/ postposé favoriserait l'antériorisation de /ɔ/, à la différence de /p/ ou /f/ par exemple ; l'étiquetage des consonnes adjacentes selon leur lieu d'articulation ne permet pas de confirmer ou d'infirmer un tel effet. Par ailleurs, comme illustré dans la FIGURE 3, les modes conditionnels de l'effet aléatoire MOT, qui rendent compte de l'impact de chaque mot sur la valeur moyenne prédite de $F_2$, révèlent des ajustements inférieurs à -100 Hz pour les 4 mots débutant par /ʁ/ (*roches*, *robes*, *rote*, *roc*). Cette observation *a posteriori* soulève la question d'un éventuel effet d'un /ʁ/ antéposé. En position postposée, l'effet de /ʁ/ n'est pas traité dans le cadre de cette contribution pour deux raisons : a) en français hexagonal, il reste controversé (Coveney, 2001 ; Armstrong et Low, 2008) et b) en français québécois, le /ʁ/ postposé en syllabe fermée induit une diphtongaison plutôt qu'une antériorisation de /ɔ/ (Arnaud et Riverin-Coutlée, 2016). Il reste que les résultats proposés dans la présente contribution indiquent que les consonnes antéposées, comme les consonnes postposées, influencent l'antériorisation de la voyelle

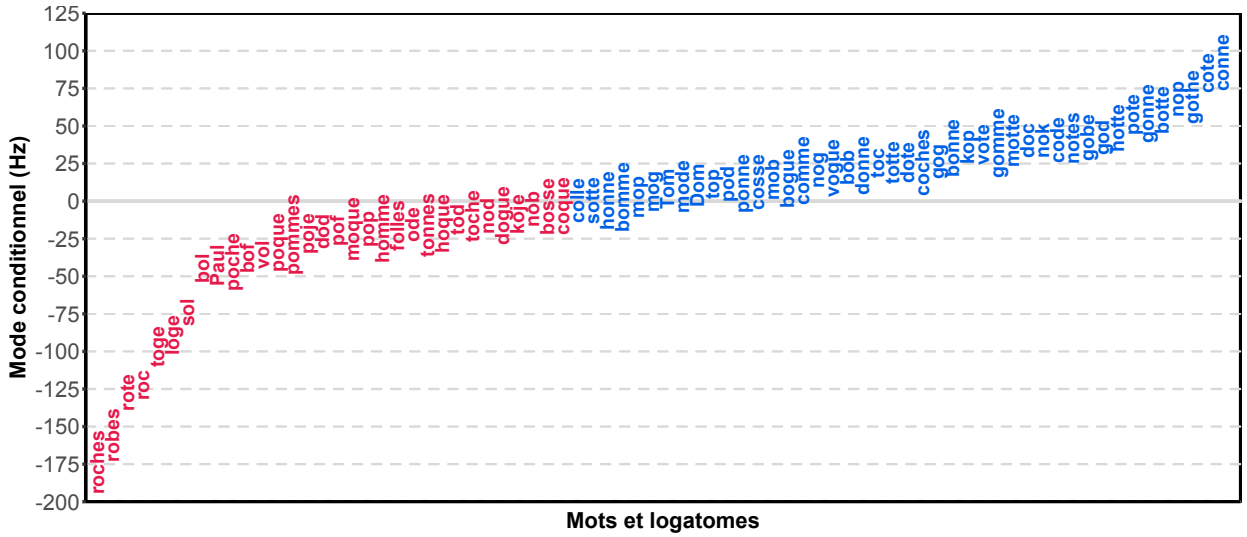



FIGURE 3 – Modes conditionnels de l'effet aléatoire MOT par ordre croissant

En conclusion, certaines interrogations demeurent. Au plan acoustique, la différence prédite pour $F_2$ entre le /ɔ/ de Saguenay et de Lyon, sexes et lieux d'articulation des consonnes adjacentes confondus, est de 137 Hz. Cette différence substantielle est-elle perçue de la même manière de chaque côté de l'Atlantique ? Des études perceptives devront être menées pour explorer cette question. Par ailleurs, l'antériorisation de /ɔ/ reste à être examinée en synchronie à l'échelle d'autres régions françaises et québécoises, notamment dans les très grands centres urbains socialement hétérogènes que sont Québec et Montréal.

# Remerciements

Cette recherche a été rendue possible grâce au soutien financier du Conseil de recherche en sciences humaines du Canada (CRSH), du Fonds de recherche du Québec – Société et culture (FRQ-SC), du Laboratoire Dynamique du Langage (UMR 5596 – CNRS Université Lyon 2) et du LABEX ASLAN (ANR-10-LABX-0081) de l'Université de Lyon (ANR-11-IDEX-0007).

# Références

ARMSTRONG N., LOW J. (2008). C'est encœur plus jeuli, le Mareuc: Some evidence for the spread of /ɔ/-fronting in French. *Transactions of the Philological Society* 106(3), 432-455.

ARNAUD V., RIVERIN-COUTLÉE J. (2016). De l'acoustique à la perception : la confusion des voyelles /ɑ/ et /ɔ/ en syllabe fermée par /ʁ/ en français québécois. *Association of French Language Studies Conference 2016*, Queen's University, Belfast.

BARR D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology* 4, 328.

BATES D., MÄCHLER M., BOLKER B., WALKER S. (2015). Fitting linear mixed-effects models using *lme4*. *Journal of Statistical Software* 67, 1-48.

CARTON F. (2000). La prononciation. *In* Cerquiglini B., Antoine G. (éds.) *Histoire de la langue française 1945-2000*. Paris : CNRS, 25-60.

COVENEY A. (2001). *The Sounds of Contemporary French: Articulation and Diversity*. Exeter : Elm Bank Publications.

DRAGER K., HAY J. (2012). Exploiting random intercepts: Two case studies in sociophonetics. *Language Variation and Change* 24, 59-78.

FONAGY I. (1989). Le français change de visage ? *Revue romane* 24(2), 225-253.

HANSEN A. B., JUILLARD C. (2011). La phonologie parisienne à trente ans d'intervalle – Les voyelles à double timbre. *Journal of French Language Studies* 21, 313-359.

HURLBERT S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54 (2), 187-211.

LAMONTAGNE J. (2015). A little forward in Laurentian French: A variationist analysis of vowel fronting in Laurentian French. *Congrès de l'Association canadienne de linguistique 2015*, 1-14.

LENNIG M. (1979). Une étude quantitative du changement linguistique dans le système vocalique parisien. *In* Lennig M., Thibault P. (éds.). *Le français parlé : études sociolinguistiques*. Edmonton : Edmonton Linguistic Research, 29-39.

MALDEREZ I. (1995). *Contribution à la synchronie dynamique du français : le cas des voyelles orales arrondies (perception et production)*. Thèse de doctorat : Université Paris VII.

MARTINET A. (1957). "C'est jeuli, le Mareuc !". *Romance Philology* 11, 345-355.

MOONEY D. (2016). 'C'est jeuli, la Gasceugne!': l'antériorisation du phonème /ɔ/ dans le français régional du Béarn. *French Studies: A Quarterly Review* 70, 61-81.

MURTAUGH P.A. (2014). In defense of *P* values. *Ecology* 95, 611-617.

NAKAGAWA S., SCHIELZETH H. (2013). A general and simple method for obtaining $R^2$ from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4(2), 133-142.

PARADIS C. (1985). *An Acoustic Study of Variation and Change in the Vowel System of Chicoutimi-Jonquiere (Quebec)*. Thèse de doctorat : University of Pennsylvania.

PINHEIRO J., BATES D. (2006). *Mixed-effects models in S and S-PLUS*. Berlin : Springer.

WALTER H. (1976). *La dynamique des phonèmes dans le lexique du français contemporain*. Paris : Presses universitaires de France.

WOEHRLING C. (2009). *Accents régionaux en français : perception, analyse et modélisation à partir de grands corpus*. Thèse de doctorat : Université Paris-Sud XI.

WOEHRLING C., BOULA DE MAREÜIL P. (2007). Comparing Praat and Snack formant measurements on two large corpora of northern and southern French. *Interspeech 2007*, 1006-1009.

ZUUR A.F., IENO E.N., WALKER N.J., SAVELIEV A.A., SMITH G.M. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Berlin : Springer.

# Étude acoustique de la production de voyelles de l'anglais par des apprenants francophones

Jennifer Krzonowski[1] François Pellegrino[1] Emmanuel Ferragne[2]
(1) Laboratoire Dynamique du Langage, 14 avenue Berthelot, 69005 Lyon, France
(2) CLILLAC-ARP EA 3967 / Université Paris Diderot
jennifer.krzonowski@cnrs.fr, françois.pellegrino@cnrs.fr,
emmanuel.ferragne@univ-paris-diderot.fr

## RESUME

Les apprenants francophones tardifs de l'anglais présentent un accent étranger dont la littérature souligne qu'il est particulièrement marqué, en ce qui concerne les voyelles, pour /ɪ/, /ʌ/ et /æ/ qui seraient prononcées « à la française ». Nous avons étudié les paramètres acoustiques des voyelles /ɪ/, /i:/, /æ/, /ʌ/ et /ɑ:/ de l'anglais produites par 38 locuteurs natifs de l'anglais et par 48 apprenants de langue maternelle française, ainsi que ceux des voyelles /a/, /e/, /i/ et /œ/ du français produites par ces mêmes apprenants. Les résultats montrent tout d'abord une plus grande variabilité des productions en L2 qu'en L1. De plus, les productions en L2 du contraste /i: - ɪ/ présentent une grande confusion entre les deux catégories, alors que l'effet inverse est observé dans la région du [a]. Nos données montrent également que les apprenants produisent les contrastes de durée mais de manière moins marquée que les locuteurs natifs. Enfin, l'étude de la dynamique des formants pour les voyelles /i:/ et /i/ suppose une variabilité dans la diphtongaison du /i:/ produit par les natifs.

## ABSTRACT

**An acoustic study of English vowels produced by French learners.**

Like every late L2-learner, French native speakers are characterized by a foreign accent. The literature emphasizes that the vowels /ɪ/, /ʌ/ and /æ/ in particular are pronounced "à la française". This study presents the acoustic parameters of the English vowels /ɪ/, /i:/, /æ/, /ʌ/ and /ɑ:/ produced by 38 native English speakers and 48 French learners of English compared to those of the French vowels /a/, /e/, /i/ et /œ/ produced by the same learners. The results first show more variability in the L2 compared to L1 productions. Furthermore, we observe a great confusion between categories in L2 for the /i: - ɪ/ contrast and the opposite effect in the region of the [a]. We show that L2 learners did produce duration contrasts but not to the same extent as the natives. Finally, the formant dynamic analysis of /i:/ and /i/ may suggest that the English native speakers do not all produce a diphthongal /i:/.

MOTS-CLES : acquisition d'une langue seconde, phonologie, voyelle, anglais langue seconde
KEYWORDS: second language acquisition, phonology, vowel, English as an L2


10.21437/JEP.2018-60

# 1 Introduction

Les apprenants de L2 tardifs sont souvent reconnaissables à leur accent étranger (Munro, 2008). Ce phénomène est bien décrit pour différentes combinaisons L1-L2 et pour différents contrastes phonologiques. L'explication théorique dominante attribue ces difficultés de production à un biais dans la perception de la L2 lié à la phonologie de la L1 et de ses liens avec la L2. Les modèles théoriques considèrent que l'acquisition de la production de la L2 suit celle de la perception. Une fois que les catégories de la L2 sont établies en perception, elles sont utilisées pour guider la production (Best & Tyler, 2007; Flege, 1995; Schwartz & Sprouse, 1996).

Dans l'enseignement des langues étrangères en France, la phonétique n'est que rarement abordée explicitement. Il a d'ailleurs été observé que des professeurs stagiaires présentaient de nombreuses difficultés de prononciation de l'anglais (Voise, 2010). En ce qui concerne les voyelles, cette étude rapporte notamment que plusieurs voyelles sont produites « à la française ». En effet, le /ɪ/ anglais serait produit comme un /i/ français, le /ʌ/ comme un /œ/ et le /æ/ comme un /a/.

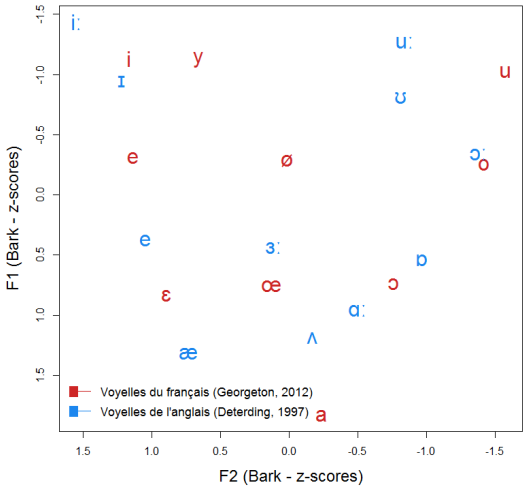



FIGURE 1 : Voyelles du français (d'après Georgeton et al., 2012, en rouge) et monophtongues de l'anglais (d'après Deterding, 1997, en bleu), pour les femmes.

La comparaison des systèmes vocaliques (monophtongues) des deux langues (FIGURE 1), montre que dans la région du [a], une seule catégorie existe en français alors que l'anglais en présente trois, /æ/, /ʌ/ et /ɑ:/, qui se situent entre le /a/ français et le /œ/. Dans la région du [i], une seule catégorie est présente en français, alors que deux catégories occupent cette région en anglais, le /i:/ et le /ɪ/. Outre le timbre des voyelles, la durée a une valeur phonologique en anglais, contrairement au français. De plus, ce contraste de durée produit en anglais une légère diphtongaison de certaines voyelles, et notamment du /i:/ (Collins & Mees, 2013; Ferragne, 2008). Nous nous sommes donc intéressés, du point de vue acoustique, à la manière dont des apprenants francophones produisent ces voyelles et leurs voisines à la fois dans l'espace vocalique anglais et français. Deux approches ont été utilisées pour décrire des productions des voyelles de l'anglais en L1 et en L2, et du français en L1 : une approche statique pour l'ensemble des voyelles et une approche dynamique pour caractériser les voyelles /i:/ produites par des locuteurs natifs et des apprenants et la voyelle /i/ produite par les locuteurs francophones.

# 2 Méthodologie

## 2.1 Procédure expérimentale

48 participants de langue maternelle française inscrits en 1ère année de LEA ou LLCER anglais (dont 16 hommes) et 38 participants (dont 15 hommes) de langue maternelle anglaise originaires du Sud-Est de l'Angleterre ont participé à l'étude. Ils ont été enregistrés lors de tâches de lecture de mots isolées (Iverson, Pinet, & Evans, 2012). Tous ont enregistré trois occurrences des voyelles anglaises /ɪ/, /i:/, /æ/, /ʌ/ et /ɑ:/ en contexte bVt (*bit, beat, bat, but, bart*[1]). Les locuteurs français ont également enregistré trois occurrences des voyelles françaises /a/, /e/, /i/ et /œ/ en contexte bV(R) (*bas, bée, bi, beurre*) dans une tâche à part. Les enregistrements ont été réalisés avec le logiciel ROCme! au format PCM mono 44 kHz 16 bits.

## 2.2 Analyse des données

Des analyses acoustiques ont été réalisées avec le logiciel Praat (Boersma & Weenink, 2017). Les voyelles ont été segmentées manuellement puis les trois premiers formants ont été mesurés semi-automatiquement sur toute la durée de la voyelle : l'estimation formantique de Praat superposée au spectrogramme était ajustée jusqu'à ce que l'estimation soit cohérente avec le spectrogramme. Cinq occurrences ont été retirées des analyses car les formants n'étaient pas visibles dans les spectrogrammes. Au total, le corpus comporte 1961 voyelles avec entre 113 et 142 occurrences par voyelles et par groupe. Les valeurs des deux premiers formants ont été extraites au milieu temporel des voyelles, transformées en Bark (Traunmüller, 1990) puis centrées-réduites par locuteur indépendamment pour chaque formant. Les durées des voyelles ont également été extraites.

# 3 Résultats

## 3.1 Analyse statique des formants

Les analyses du timbre des voyelles ont été réalisées séparément pour les hommes et les femmes. Dans les représentations graphiques qui suivent, les données des femmes sont représentées à gauche, celles des hommes à droite. Pour chacun des graphiques, les symboles phonétiques indiquent le barycentre de la catégorie, et les ellipses représentent l'intervalle de confiance au seuil de 95% des valeurs moyennes des formants pour les groupes de participants concernés. Des proportions de chevauchement entre les catégories voisines ont été mesurées en faisant le rapport entre l'aire de l'intersection par l'aire de l'union des deux catégories considérées.

[1] L'anglais du Sud-Est de l'Angleterre est proche de la variété d'anglais enseignée en France, d'où cette restriction dans les critères d'inclusion. Il faut noter que dans cette variété, le r en coda ne se prononce pas. Ainsi *bart* se prononce [bɑ:t].

### *3.1.1 Comparaison des voyelles anglaises et françaises produites par des natifs*

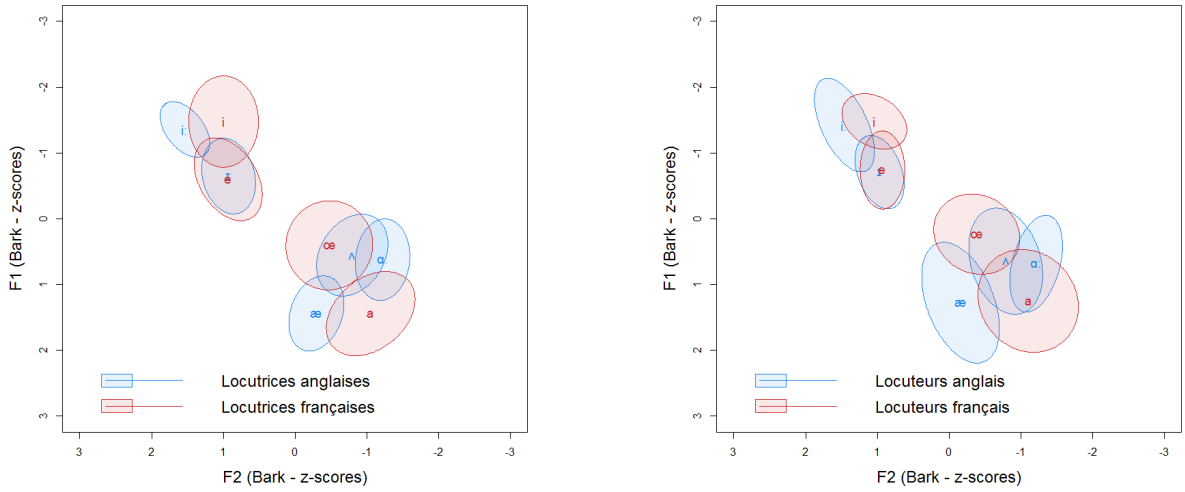



FIGURE 2 : Voyelles de l'anglais (en bleu) et du français (en rouge) produites par des locuteurs natifs, pour les femmes à gauche, pour les hommes à droite.

Une analyse de la variance sur les aires des ellipses pour chaque catégorie vocalique ne montre pas d'effet de la Langue, du Sexe des participants, ni d'interaction Langue × Sexe. Pour les femmes et les hommes, on constate que les aires des ellipses sont équivalentes en anglais et en français ($F_{(1, 14)} = 2.06$, $p > .05$). Il semble donc que la variabilité intra-catégorielle produite par des locuteurs natifs soit équivalente d'une langue à l'autre.

Dans la région acoustique du [i], le /i/ français occupe un espace distinct des catégories anglaises : il est moins antérieur que le /iː/ anglais et plus fermé que /ɪ/. Ainsi il partage peu d'espace acoustique avec ces catégories : 12% de chevauchement chez les femmes, 4% chez les hommes pour /ɪ/ ; 9% chez les femmes et 15% chez les hommes pour /iː/. En revanche, le /e/ français occupe la même zone que le /ɪ/ anglais avec 72% de superposition entre les deux catégories pour les femmes, 75% pour les hommes. On peut ainsi s'attendre à ce que les locuteurs français produisent cette voyelle avec les routines articulatoires du /e/ français.

Dans la région du [a], pour les hommes et les femmes, le /æ/ occupe une zone plus antérieure que les autres voyelles françaises et anglaises et ne partage ainsi que très peu son espace acoustique (5% avec /a/, 3% avec /ʌ/ et 3% /œ/ chez les femmes et 4% avec /a/, 2% avec /ʌ/ et 5% /œ/ chez les hommes). Il en est de même pour le /a/ français chez les femmes, qui est plus ouvert que les voyelles /œ/, /ʌ/ et /ɑː/, plus postérieur que le /æ/ anglais et partage peu d'espace acoustique avec elles (3% avec /ʌ/, 11% avec /ɑː/, 0% avec /œ/). Le profil est un peu différent chez les hommes puisque cette voyelle est moins ouverte et partage donc davantage d'espace avec les voyelles /ʌ/ (27%) et /ɑː/ (24%). On peut enfin noter une grande confusion dans les espaces acoustiques des voyelles /œ/ et /ʌ/ (41% de chevauchement chez les femmes, 29% chez les hommes). On peut s'attendre ici à ce que les français réemploient les schémas articulatoires de cette voyelle pour produire le /ʌ/ de l'anglais comme observé dans l'étude de Voise (2010). On remarque enfin que parmi les voyelles de L1 étudiées dans les deux langues, les voyelles de l'anglais /ʌ/ et /ɑː/ sont celles qui partagent les plus grandes proportions de leurs espaces acoustiques (23% pour les femmes, 24% pour les hommes).

### *3.1.2 Comparaison des voyelles anglaises et françaises produites par des locuteurs français*

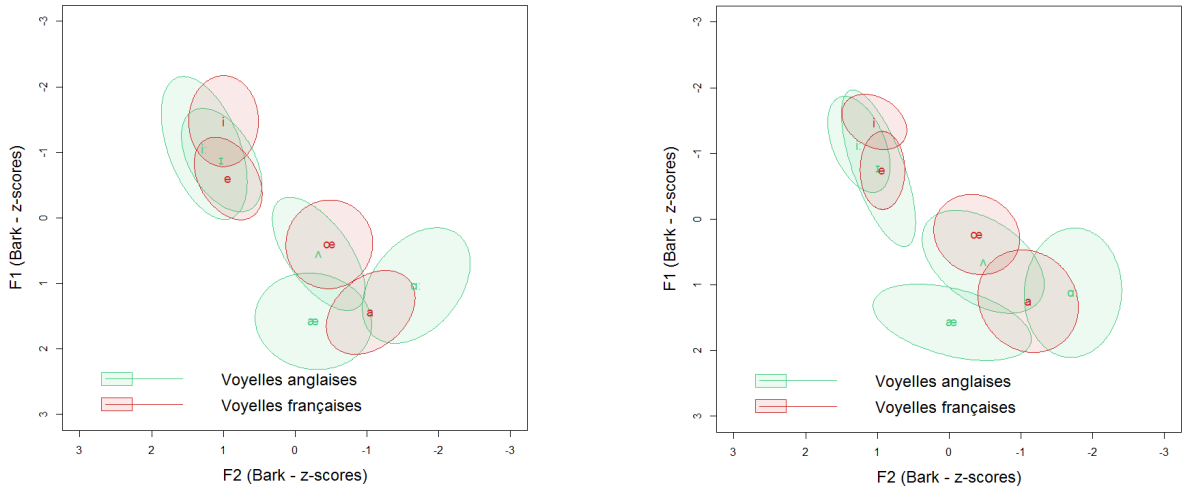



FIGURE 3 : Voyelles produites par les locuteurs français, en anglais (en vert) et en français (en rouge), pour les femmes à gauche, pour les hommes à droite.

Une analyse de la variance sur les aires des ellipses pour chaque catégorie vocalique montre cette fois un effet du facteur Langue ($F_{(1, 14)} = 9.40$ ; $p < .01$), indiquant une plus grande variabilité des productions des locuteurs français lorsqu'ils produisent des sons de L2 que de L1 et conduisant à de forts chevauchements entre les différents espaces des voyelles.

Dans la région acoustique du [i], on remarque une forte confusion dans les espaces des voyelles anglaises /i:/ et /ɪ/ (50% pour les femmes, 41% pour les femmes) probablement due à la grande variabilité des productions. Néanmoins, on observe que la configuration des centres de catégories par rapport aux voyelles françaises semble tendre vers les cibles anglaises. En effet, le /i:/ est plus postérieur et plus ouvert que le /i/ français, le /ɪ/ plus ouvert que le /i/ et se rapprochant ainsi du /e/.

Dans la région du [a], il est frappant de constater que les trois catégories anglaises sont produites dans des espaces acoustiques distincts avec peu de chevauchement entre les catégories (< 10%). Comme attendu, on observe une grande confusion entre les espaces du /œ/ et du /ʌ/ (52% pour les femmes, 40% pour les hommes). L'espace acoustique du /a/ chevauche les trois catégories de l'anglais mais les centres des catégories restent bien distincts.

### *3.1.3 Comparaison des voyelles anglais produites en L1 et L2*

Une analyse de la variance réalisée sur les aires des ellipses indique ici encore que les productions des français en L2 présentent une plus forte variabilité que celles des anglais en L1 et ce pour toutes les voyelles ($F_{(1, 16)} = 69.31$, $p < .001$).

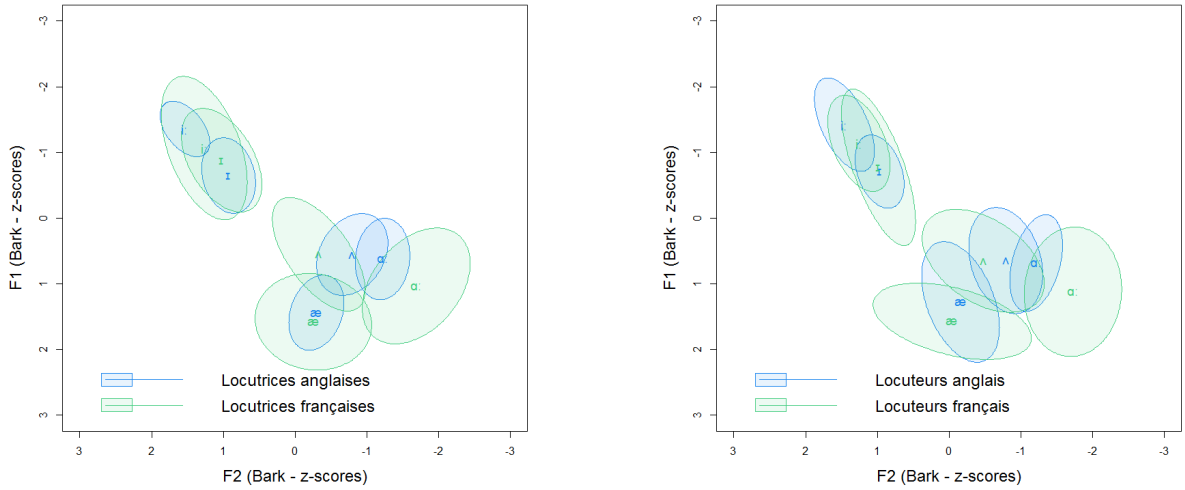



FIGURE 4: Voyelles anglaises produites par des locuteurs natifs (en bleu) ou de L2 (en vert), à gauche pour les femmes, à droite pour les hommes.

Dans la région du [i], on observe que les voyelles produites par les natifs sont plus distinctes que celles produites par les apprenants (0% de chevauchement entre les deux catégories chez les femmes, 8% chez les hommes pour les natifs versus 50% pour les femmes, 41% pour les hommes chez les apprenants). De plus, on observe une plus grande distance acoustique entre les centres des catégories des natifs (M = 2.055) que ceux des apprenants (M = 1.94).

Dans la région du [a], on remarque le phénomène inverse : les voyelles des apprenants sont plus distinctes que celles des locuteurs natifs. On observe un chevauchement moins important entre les trois catégories produites par les locuteurs francophones que par les locuteurs natifs, de plus, les distances acoustiques dans F1/F2 entre les centres des catégories des francophones sont plus grandes que celles des locuteurs natifs. Il semble que dans cette région, les participants français utilisent davantage le timbre des voyelles pour produire des voyelles qui soient à la fois distinctes entre elles mais aussi pour le /æ/ et le /ɑ:/ différentes du /a/ français. Au sujet du /ʌ/, on note que les productions des apprenants s'étendent sur une région bien plus centrale que les productions des natifs, ce qui vient appuyer l'idée que les apprenants utilisent le /œ/ comme cible articulatoire.

## 3.2 Analyse des durées des voyelles

Les voyelles de l'anglais présentent des différences de durées significatives ($F_{(4, 148)} = 258.57$, $p < .0001$). Les voyelles /æ/ (290 ms), /ɑ:/ (340 ms) et /i:/ (287 ms) sont plus longues que les voyelles /ʌ/ (172 ms) et /ɪ/ (159 ms) (FIGURE 5, en bleu). Les voyelles du français présentent également des différences de durées significatives ($F_{(3, 141)} = 55.80$, $p < .001$), mais cette différence n'est portée que par la voyelle /œ/ produite en contexte bVR, qui a pour effet d'allonger la durée de la voyelle (Léon, 2005) (FIGURE 5, en rouge). Ainsi les voyelles du français ont des durées similaires aux voyelles courtes de l'anglais (sauf le /œ/ qui est plus long) et toutes sont significativement plus courtes que les voyelles longues de l'anglais (y compris /œ/).

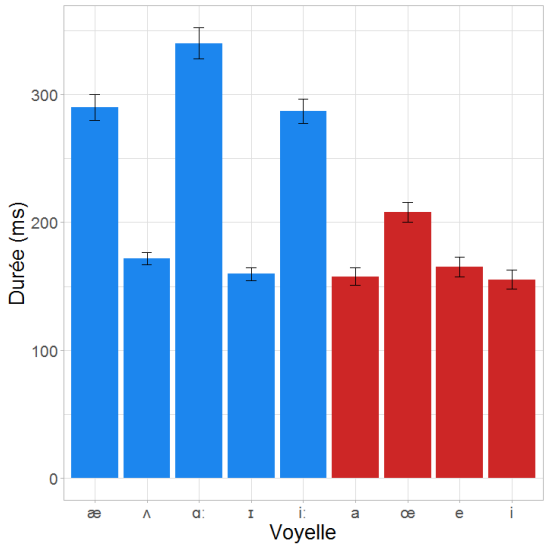



FIGURE 5 : Durées moyennes des voyelles anglaises (en bleu) et françaises (en rouge) produites par des locuteurs de L1.

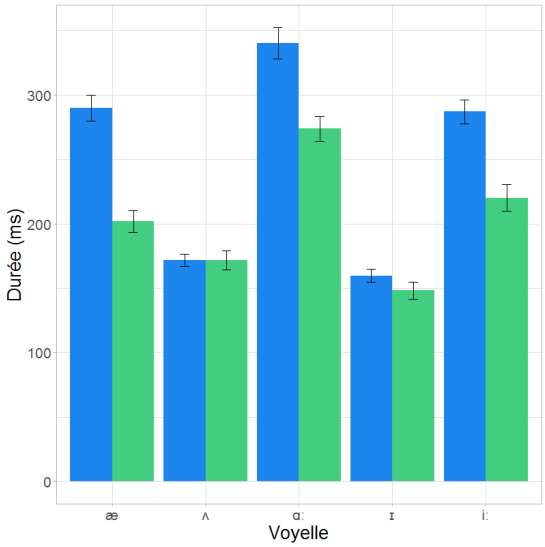



FIGURE 6 : Durées moyennes des voyelles anglaises produites par des locuteurs de L1 (en bleu) et de L2 (en vert).

Lorsque les apprenants français produisent les voyelles anglaises (FIGURE 6, en vert), les durées des voyelles courtes (/ʌ/ et /ɪ/) sont similaires aux durées de ces mêmes voyelles produites par des locuteurs natifs (en bleu) ($F_{(1, 84)} = 1.66, p > .05$). En revanche, pour les voyelles longues, on observe que les durées des voyelles produites par les locuteurs français sont significativement plus courtes que celles produites par les locuteurs natifs ($F_{(1, 84)} = 7.732, p < .001$). Cependant, pour ces mêmes voyelles, on observe bien un allongement de la durée par rapport aux voyelles françaises ($F_{(1, 335)} = 14.56, p < .001$). Ainsi, il semblerait qu'en L2, les apprenants français soient capables de produire un contraste de durée bien qu'il n'en existe pas dans leur L1.

## 3.3 Analyse dynamique des voyelles /i:/ et /i/

Suivant Williams & Escudero (2014), les 80% centraux de chaque courbe de F2 des voyelles /i :/ et /i/ produites par des natifs et des voyelles /i:/ produites par les apprenants ont été ré-échantillonnées afin d'avoir 30 points. Les coefficients 2 et 3 (DCT2 et DCT3) d'une transformée en cosinus discrète ont été calculés. Ces coefficients permettent de caractériser les trajectoires de formants. Les comparaisons de différents modèles d'analyse discriminante linéaire montrent que la valeur des deux premiers formants seuls prise au milieu temporel suffit à distinguer les voyelles /i:/ (97%) et /i/ (90%) produites par des natifs lorsque ces seules voyelles sont intégrées dans l'analyse. Un modèle incluant uniquement DCT2 classifie correctement les /i/ français (87%) mais pas les /i:/ anglais (46%), ce qui indique que ce paramètre est plus pertinent pour classifier les /i/ français que les /i:/ anglais, probablement à cause de la non-systématicité de la diphtongaison du /i:/ anglais. Le meilleur modèle pour classifier ces voyelles natives a été ensuite utilisé pour classifier les /i:/ des apprenants. Ce modèle porte sur F1, F2, la durée et DCT2. Ce modèle classe 58% des voyelles de L2 comme des /i:/ natif et 42% comme des /i/ français. Bien que non significatif, ($\chi^2_{(1)} = 3.36, p = .067$), ce résultat semble indiquer qu'au moins une partie des /i:/ produits par les apprenants présentent des paramètres acoustiques semblables à ceux des voyelles produites par les natifs, y compris des paramètres dynamiques.

# 4 Discussion

Pour les voyelles étudiées, la comparaison des espaces vocaliques – anglais natif, français natif et anglais L2 – montre tout d'abord une plus grande variabilité des voyelles produites en L2. Ceci peut s'expliquer par le fait qu'en L2 les cibles articulatoires ne sont pas clairement établies chez les apprenants (Gick, Bernhardt, Bacsfalvi, & Wilson, 2008).

Pour les contrastes de l'anglais /i: - ɪ/ et /ʌ - ɑ:/, qui portent à la fois sur le timbre et la durée, nous avons observé des phénomènes différents dans les productions des locuteurs natifs. En effet, les catégories /i:/ et /ɪ/ sont produites par les natifs avec une plus grande distance acoustique entre les centres de catégories que ne le sont les voyelles /ʌ/ et /ɑ:/. Le plus grand chevauchement des aires des catégories /ʌ/ et /ɑ:/ résultant est contrebalancé par un contraste de durée plus marqué pour ce contraste. Chez les apprenants, ce schéma est différent : ils produisent le contraste /i: - ɪ/ avec une faible distance acoustique entre les catégories et donc un fort chevauchement entre leurs aires, alors qu'ils produisent le contraste /ʌ - ɑ:/ avec une grande distance entre les catégories et un faible chevauchement entre leurs aires, et ce, malgré une grande variabilité intra-catégorielle. Il semblerait que, pour les voyelles de la région du [a] (/æ/ compris), les apprenants cherchent à produire des voyelles qui se distinguent du point de vue du timbre, à la fois entre elles, mais aussi du /a/ français. Cette observation renvoie au phénomène de dissimilation décrit dans le *Speech Language Model* de Flege (1995). Selon ce modèle, la probabilité de formation d'une catégorie phonétique augmente avec la dissimilarité entre un son de L2 et le plus proche voisin en L1. En l'absence de formation de nouvelle catégorie, les propriétés d'un son L2 et d'un son L1 voisin peuvent fusionner pour former une catégorie « composite ». Les éléments phonétiques des deux inventaires coexistent dans un espace phonologique commun et interagissent les uns avec les autres conduisant parfois à des déviations (assimilation ou dissimilation) des catégories voisines de L1, L2 et L1/L2 pour maintenir les contrastes phonétiques.

Pour le contraste /i: - ɪ/, les apprenants semblent au contraire présenter des difficultés à produire des voyelles distinctes du point de vue du timbre, ne produisant quasiment qu'un contraste de durée. Bien que des précautions aient été prises lors de la présentation des consignes (le mot était accompagné du symbole phonétique de la voyelle), on peut imaginer pour ce contraste que les participants francophones aient été influencés par le système orthographique de leur L1 selon lequel la graphie du mot « *bit* » induit la production d'un [i] (Nimz, 2016). Selon cette interprétation, la réalisation du /ɪ/ tendrait à se rapprocher de celle du /i/ pour des raisons orthographiques, et ce malgré le fait que le /e/ soit acoustiquement un meilleur support articulatoire pour la réalisation du /ɪ/. Au sujet de la voyelle /ʌ/ plus particulièrement, on a pu remarquer que la distribution des voyelles produites par les apprenants s'étendait de manière plus importante que chez les natifs en direction de la région du /œ/, ce qui confirme que cette voyelle pourrait en effet servir de support articulatoire aux apprenants francophones.

En ce qui concerne la durée des voyelles, nos données montrent que, bien que la durée ne soit pas contrastive en français, les apprenants francophones produisent des voyelles longues de l'anglais plus longues que les voyelles françaises même si celles-ci restent moins longues que celles des locuteurs natifs.

L'analyse dynamique pour les voyelles /i:/ de l'anglais et /i/ du français montre tout d'abord que les indices statiques des deux premiers formants seuls suffisent à catégoriser ces voyelles alors

qu'un indice dynamique seul n'est pertinent que pour les voyelles françaises, ce qui pourrait indiquer une certaine variabilité dans les productions des /i:/ par les natifs (Ferragne, 2008). Enfin, une partie des productions des apprenants présentent des paramètres similaires aux voyelles natives.

# Remerciements

Ce travail est soutenu par la subvention de recherche de l'IUF d'E. Ferragne et le LabEx ASLAN de l'Université de Lyon (ANR-10-LABX-0081).

# Références

BEST, C. T., & TYLER, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. *Language experience in second language speech learning: In honor of James Emil Flege*, *1334*.

BOERSMA, P., & WEENINK, D. Praat: doing phonetics by computer. (Version 6.0.26).

COLLINS, B., & MEES, I. M. (2013). *Practical Phonetics and Phonology: A Resource Book for Students*. Routledge.

DETERDING, D. (1997). The Formants of Monophthong Vowels in Standard Southern British English Pronunciation. *Journal of the International Phonetic Association*, *27*(1‑2), 47‑55.

FERRAGNE, E. (2008). Etude Phonétique des Dialectes Modernes de l'Anglais des Iles Britanniques: Vers l'Identification Automatique du Dialecte. *Unpublished doctoral thesis, Université Lyon*, *2*.

FLEGE, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, *92*, 233–277.

GEORGETON, L., PAILLEREAU, N., LANDRON, S., GAO, J., & KAMIYAMA, T. (2012). Analyse formantique des voyelles orales du français en contexte isolé : à la recherche d'une référence pour les apprenants de FLE (p. 145‑152). Présenté à Conférence conjointe JEP-TALN-RECITAL 2012.

GICK, B., BERNHARDT, B., BACSFALVI, P., & WILSON, I. (2008). 11. Ultrasound imaging applications in second language acquisition. In J. G. Hansen Edwards & M. L. Zampini (Éd.), *Studies in Bilingualism* (Vol. 36, p. 309‑322). Amsterdam: John Benjamins Publishing Company.

IVERSON, P., PINET, M., & EVANS, B. G. (2012). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics*, *33*(1), 145‑160.

LÉON, P. R. (2005). *Phonétisme et prononciations du français*. A. Colin.

MUNRO, M. J. (2008). 7. Foreign accent and speech intelligibility. In J. G. Hansen Edwards & M. L. Zampini (Éd.), *Studies in Bilingualism* (Vol. 36, p. 193‑218). Amsterdam: John Benjamins Publishing Company.

NIMZ, K. (2016). Sound perception and production in a foreign language.

SCHWARTZ, B. D., & SPROUSE, R. A. (1996). L2 cognitive states and the Full Transfer/Full Access model. *Second language research*, *12*(1), 40–72.

TRAUNMÜLLER, H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, *88*(1), 97–100.

VOISE, A.-M. (2010). Enseigner la phonologie de l'anglais aux futurs professeurs du primaire. *Recherche et pratiques pédagogiques en langues de spécialité. Cahiers de l'Apliut*, (Vol. XXIX N° 2), 11‑24.

WILLIAMS, D., & ESCUDERO, P. (2014). A cross-dialectal acoustic comparison of vowels in Northern and Southern British English. *The Journal of the acoustical society of America*, *136*(5), 2751–2761.

XXXIIe Journées d'Études sur la Parole
4-8 juin 2018, Aix-en-Provence, France

# Étude acoustique du cluster /t͡ɾ/ et de ses allophones à Santiago du Chili

Alexis Dehais Underdown[1] Didier Demolin[1]

(1) Laboratoire de Phonétique et Phonologie, (CNRS / Sorbonne Nouvelle), 19 Rue des Bernardins, 75005 Paris, France

alexis.underdown@gmail.com, didier.demolin@univ-paris3.fr

## RESUME

Cette étude a pour but d'étudier la production du groupe allophonique /t͡ɾ/ à Santiago du Chili afin de donner les caractéristiques acoustiques des rhotiques. Pour cela, un protocole d'élicitation de parole (jeu de dénomination d'images) a été créé. Les paramètres prosodiques ne sont pas traités ici et le débit de parole n'a pas été contrôlé durant l'expérience. Les données montrent différentes réalisations du cluster (comme [t̪ᵊɾ̪] [t̪͡s̪] ou [t̪̊]) ainsi que l'insertion d'un vocoïde central [ᵊ] dans les réalisations non-affriquée et non-fricative. Cette insertion paraît empêcher les processus de coarticulation entre les consonnes coronales /t̪/ et /ɾ/. De plus, nous pensons que ce vocoïde pourrait jouer un rôle dans la perception du cluster, en particulier de celle du tap apico-alvéolaire occlusif qui suit. Nos données montrent que les allophones [t̪͡s̪̊] et [t̪̊] ne sont pas motivés par des facteurs linguistiques mais plutôt par des facteurs sociolinguistiques comme l'âge ou le profil socio-économique.

## ABSTRACT

**Allophonic variation of /t͡ɾ/ cluster production in Santiago de Chile.**

This paper aims to highlight the production of /t͡ɾ/ allophonic group in Santiago de Chile in order to provide acoustic characteristics of the rotics. For this we created an elicitation task protocol (picture-naming game). Prosodic parameters were not analyzed and speech rate was not controlled during the experimentation. Results showed different realizations of the cluster (such as [t̪ᵊɾ̪] [t̪͡s̪̊] or [t̪̊]) as well as a central vocalic intrusion [ᵊ] in the non-affricated and non-fricative productions. This insertion seems to prevent coarticulation effects between /t̪/ and /ɾ/ coronal consonants. Furthermore, we think that the vocoïde may play a role in the cluster's perception in particular of the following apico-alveolar tap stop. Data shows that [t̪͡s̪̊] and [t̪̊] are not motivated by linguistic factors but rather by sociolinguistic factors as age or socioeconomic profile.

MOTS-CLES : espagnol chilien, cluster consonantique, variation allophonique, intrusion vocalique

KEYWORDS: Chilean Spanish, consonantal cluster, allophonic variation, vocalic intrusion

## 1. Introduction

Le présent travail porte sur la variation allophonique du cluster /t͡ɾ/ dans l'espagnol parlé au Chili. Notre attention portera sur la variation, notion centrale en linguistique. La variation peut être motivée par des facteurs divers : linguistiques, pathologiques, émotionnels, extralinguistiques (identité, sexe, âge etc…). Nous étudierons plus précisément la variation phonétique dudit cluster pour comprendre ses caractéristiques ainsi que celui de ses allophones. /t͡ɾ/ n'est pas un phonème à proprement parler, mais on utilisera la notation phonologique, comme le suggère Sadowsky (2015), pour indiquer qu'il s'agit de la forme phonologique et de l'allophone non marqué de son


10.21437/JEP.2018-61

groupe. Ce groupe est un continuum d'allophones allant du cluster [t͡ɾ̪] aux allophones fricatifs ([ɹ̝̊]), en passant par des allophones affriqués ([t͡ɹ̝̊]). Cette allophonie n'est pas propre au parlé chilien, selon Alonso (1953) on retrouve les réalisations affriquées (comparable au phone affriqué [t͡ɹ̝̊] de l'anglais [[t͡ɹ̝̊i:] <*tree*>) dans différents pays hispanophones (Guatemala, Costa Rica, Colombie, Mexique, Pérou, Argentine, Espagne etc.).

### 1.1 État de la question

Dans cette étude, nous avons pris soin d'observer les différentes réalisations de la rhotique en contexte tautosyllabique lors de l'articulation du cluster /t͡ɾ̪/. Bon nombre d'études soulignent la présence d'un vocoïde de type intrusif (non épenthétique) dans l'articulation de cluster de type /C°ɾ/ (Lenz 1940 [1892 93], Bradley 2002 et 2004, Hall 2006, Cicres & Blecua 2015) ; dans notre étude, nous nous sommes donc interrogés sur le rôle de ce vocoïde intrusif (aussi appelé élément « *svarabhatkti* », voyelle « intrusive » ou parasite). Nancy Hall (2006) étudie ce phénomène de « voyelle » intrusive du point de vue phonologique et suggère le possible rôle perceptif de celle-ci, spécialement dans les contextes tautosyllabiques où l'on retrouve une obstruente suivie d'un tap. Le tap alvéolaire /ɾ/ (i.e occlusif apico-alvéolaire) ne se réalise pas systématiquement comme tel (Bradley 2004, Cicres & Blecua 2015) mais parfois comme un tap occlusif apico-alvéolaire sourd [ɾ̥], approximant bref [ɾ̞], long [ɹ] ou bien encore comme une fricative [ɹ̝̊]. Selon Sadowsky (2015), le cluster /t͡ɾ̪/ présente un nombre d'allophones assez variés au Chili, pouvant se réaliser comme [t͡ɾ̪] [t͡ɹ̝̊] [t͡ɹ̝̊] [t͡ɹ̝̊ʱ] [ɹ̝̊] [ɹ̝̆] [ʂ] [t͡ʂ] ; l'auteur souligne que cette allophonie n'est pas motivée par des facteurs contextuels mais plutôt sociolectaux, de plus cette allophonie apparaît dans la parole spontanée et la communication informelle (Alonso 1953, Bradley 2004). D'autres auteurs soulignaient bien avant Sadowsky qu'il existait une articulation affriquée du cluster : Lenz (1893) attribuait cette articulation au substrat mapudungun (langue indigène du sud du Chili qui compte parmi son inventaire phonémique l'affriquée rétroflexe sourde [t͡ʂ]), Amado (1953) attaquait la théorie indigéniste de Lenz en signalant que l'articulation affriquée [t͡ɹ̝̊] existait aussi dans d'autre pays hispanophones et suggère une spécificité de l'espagnol.

### 1.2 Objectifs et hypothèses

Notre objectif principal est de donner une explication à l'allophonie du groupe /t͡ɾ̪/ à Santiago du Chili. Ici, l'allophonie désigne la production des phones affriqués et fricatifs [t͡ɹ̝̊] [ɹ̝̊]. Le deuxième objectif est de caractériser acoustiquement les différentes réalisations de la rhotique en contexte tautosyllabique lors de l'articulation du cluster étudié. Enfin, nous avons étudié la question du vocoïde intrusif pour en définir les rôles et fonctions. L'insertion du vocoïde ne fait pas partie de cette allophonie car, comme nous le verrons, elle empêche la production d'allophones affriqués et fricatifs

Notre hypothèse s'aligne, d'une part, avec ce que suggère Hall (2006) : le vocoïde intrusif permet une meilleure perceptibilité du tap qui suit l'occlusive; et d'autre part, nous ajouterons aussi qu'il fonctionne comme une frontière entre l'occlusive dentale sourde /t/ et le tap apico-alvéolaire voisé /ɾ/ afin d'éviter des processus d'assimilation entre les deux gestes et ainsi éviter une affrication. Enfin, nous ajouterons que l'allophonie est motivée par des facteurs sociolinguistiques. Notre hypothèse s'appuie donc sur des considérations phonologiques (Hall, 2006) et sociophonétiques (Sadowsky, 2015).

## 2. Méthodologie

La variation allophonique du groupe /t͡ɾ̪/ dépend, en grande partie, du profil socio-économique du locuteur (âge, sexe, strate sociale) ; nous avons donc essayé de monter une expérience prenant en

compte les trois variables citées précédemment. Nous avons utilisé une méthode d'élicitation de parole : la dénomination d'image.

### 2.1 Locuteurs

Nous avons enregistré 15 locuteurs natifs de l'espagnol vivant dans la capitale chilienne (Santiago). Nous avons choisi nos participants en fonction de leur profil socio-économique en nous appuyant sur une publication de l'AIM (Asociación Chilena De Empresas De Investigación De Mercado). À partir d'un système de points, ils font une classification socio-économique de la population dite « du grand Santiago » (centre et communes alentours formant la province de Santiago). Cette classification précise du statut socioéconomique se fonde sur des variables économiques (e.g. équipement du foyer, résidence secondaire) et sociales (éducation, métier) des habitants de Santiago. L'AIM réalise principalement des enquêtes de marché et d'opinion publique. Leur étude propose cinq profils socio-économiques dans la capitale : $ABC_1$ (10% de la population), $C_2$ (20%), $C_3$ (25%), D (35%) et E (10%). Dans leur répartition, le groupe A est celui qui possède le plus de biens matériels ainsi qu'une meilleure éducation (lycée privé, université privée) et par extension des métiers mieux rémunérés et plus « prestigieux » (e.g chefs d'entreprise, diplomates) ; à l'inverse le groupe E est le plus défavorisé, ses biens matériels sont très réduits et son éducation aussi (l'éducation n'est pas gratuite au Chili). En croisant cette classification avec les cartographies de Santiago montrant la répartition des strates sociales (Beatriz Mella, 2009), nous avons pu définir le profil économique de nos participants. Afin de faciliter notre travail, nous avons réduit le nombre de profils à seulement 3 : A ($ABC_1$ dans l'AIM), B ($C_2$ $C_3$ dans l'AIM), C (D et E dans l'AIM). Parmi ces 15 locuteurs, se trouvaient 10 femmes âgées de 26 à 61 ans. Nous avons aussi recueilli les enregistrements de 5 hommes âgés entre 27 à 35 ans. Nous avons 4 femmes de profil socio-économique bas (âgées de 50 à 60 ans), les 11 autres participants sont de strates sociales plus aisées. Une seule femme de 28 ans appartient au groupe A du point de vue de ses biens et de son parcours scolaire ; cependant, lors de l'analyse des données, nous avons jugé préférable de l'intégrer au groupe B car ses productions du cluster cible étaient les mêmes que les locuteurs du groupe B ($C_2$ $C_3$ dans l'AIM). En fusionnant les groupes A et B en AB, il ne reste donc que deux groupes de locuteurs : AB et C.

### 2.2 Corpus

Nous avons créé un corpus composé de 48 mots espagnols (dont certains n'existent qu'au Chili) qui regroupait 24 mots cibles contenant le cluster à l'étude et 24 distracteurs. Le son cible (i.e /t͡ɾ/) apparaissait dans différentes positions dans le mot, il était suivi de chacune des 5 voyelles de l'espagnol (i.e [a e i o u], toutes les voyelles des mots cibles ont été analysées sauf pour celles produites en voix craquée ou soufflée), d'une diphtongue, mais aussi précédé par la consonne [s] ou [n] : tra**mpa** 'piège' / le**tre**ro 'panneau' / **tri**ste 'triste' / árbi**tro** 'arbitre' / **trutru**ca 'instrument mapuche' / e**stre**lla 'étoile' / ce**ntro** 'centre' / **trián**gulo 'triangle' / monstruo 'monstre'.

### 2.3 Déroulement

Comme nous l'avons indiqué plus haut, nous avons utilisé une méthode de dénomination d'image basée sur le jeu smartphone « *4 images 1 mot* ». Rappelons que Bradley (2004) et Alonso (1953) soulignent que les productions fricatives et affriquées du groupe /t͡ɾ/ apparaissent plus souvent en situation conversationnelle spontanée et informelle ; de plus ces réalisations ne sont pas influencées par le contexte segmental ou prosodique. C'est donc pour cela que nous avons opté pour un protocole fondé sur le jeu ; bien que le jeu ne permette pas de contrôler parfaitement tous les paramètres expérimentaux et les productions des locuteurs, il permet néanmoins d'éliciter des productions plus naturelles (par rapport à une tâche de lecture ou de répétition par exemple) dans

la mesure où les participants sont en quelque sorte « distraits » par la dynamique du jeu et sont donc plus enclin à relâcher leur prononciation et produire les allophones fricatifs et affriqués. Le jeu a été présenté sur un power point où, sur chaque diapositive, figuraient quatre images en relation avec un mot cible ou avec un distracteur (e.g. quatre images de trains différents pour le mot « *tren* » 'train'). Une fois le mot cible trouvé, il a été demandé aux locuteurs de formuler la phrase de leur choix et ceci afin de faciliter la segmentation sur *Praat* lorsqu'un mot commence par l'occlusive dentale sourde [t] (e.g. « *tren* » 'train'). Nous avons divisé notre corpus sur trois power point (48 mots/3 = 16 mots par power point) afin que le déroulement de l'expérience soit plus agréable pour les participants. Afin de faciliter le jeu, les participants étaient autorisés à penser à haute voix ou même à discuter avec l'expérimentateur (ces données ont aussi été analysées). Des indices précédaient les diapositives avec les quatre images, et la réponse apparaissait sur la diapositive suivante. Les enregistrements ont été réalisés à l'aide d'un microphone (AKG, microphone électrostatique de type cardioïde) et d'une carte son (UA-25 EX) fournis par le Laboratoire de Phonétique et Phonologie UMR 7018, CNRS / Université Sorbonne Nouvelle. Nous avons utilisé le logiciel Audacity (version 2.1.2) pour enregistrer.

### 2.4 Analyse

Nous avons pris soin de rééchantillonner les enregistrements à 16000Hz pour ensuite extraire les mots cibles en vue d'une analyse sur *Praat* (version 6.0.21). Nous avons analysé les caractéristiques acoustiques de l'élément occlusif (durée totale), du vocoïde (structure formantique et durée) puis de la rhotique. Pour en effectuer l'analyse nous avons rédigé un script sur *Praat* qui a permis de traiter : (1) les différences liées au sexe (i.e. valeurs de formants), (2) la durée totale des sons occlusifs, rhotiques et de la voyelle qui suit, (3) ainsi que la structure formantique du vocoïde et des voyelles suivantes. Rappelons que notre travail portait sur la réalisation de la rhotique, nous laisserons donc de côté le contexte prosodique. Enfin, nous discuterons de l'influence des variables de sexe, d'âge et de profil socioéconomique afin d'évlauer l'impact de ces dernières sur les réalisations de [t͡i̥] et [i̥].

## 3. Résultats

### 3.1 Caractéristiques acoustiques du vocoïde de transition

À l'aide d'un script, nous avons été en mesure d'analyser le vocoïde transitionnel (n= 367) en prenant comme point de référence les voyelles de l'espagnol, à savoir /i e a o u/ (durée, F1, F2, F3). Bien que la présence du vocoïde soit très systématique (86% sur 437 occurrences de [t͡ɾ] [t͡ɾ̥] [t͡ɹ̥] [t͡ɹ]), il arrive parfois qu'il ne se réalise pas (14% des cas).

Lors de l'analyse, nous avons fait le choix de séparer d'une part, les hommes des femmes, du fait des différences liées à la longueur du conduit vocal et des valeurs de formants et d'autre part les femmes de profil socio-économique C du fait de la durée plus importante du vocoïde [ᵊ] (29ms pour le groupe C contre 22,5ms pour le groupe AB). Les valeurs formantiques moyennes du vocoïde suggèrent en effet un élément plus central que les voyelles du système : 337 Hz (F1), 1483 Hz (F2) et 2472 Hz (F3) pour les hommes et 411 Hz (F1), 1887 Hz (F2) et 2998 Hz (F3) pour les femmes. La distribution des sons sur l'espace acoustique (figure 1) indique que le vocoïde se trouve plutôt au centre de celui-ci. À gauche, nous avons la distribution non-normalisée qui compare la différence entre les femmes ([ᵊ] F1 vers 400Hz et F2 vers 1800Hz) et les hommes ([ᵊ] F1 vers 350Hz et F2 vers 1500Hz) ; le vocoïde est central et fermé. À droite, il y a la distribution normalisée (Lobanov z-score) qui va nous permettre de comparer les valeurs formantiques entre nos locuteurs.

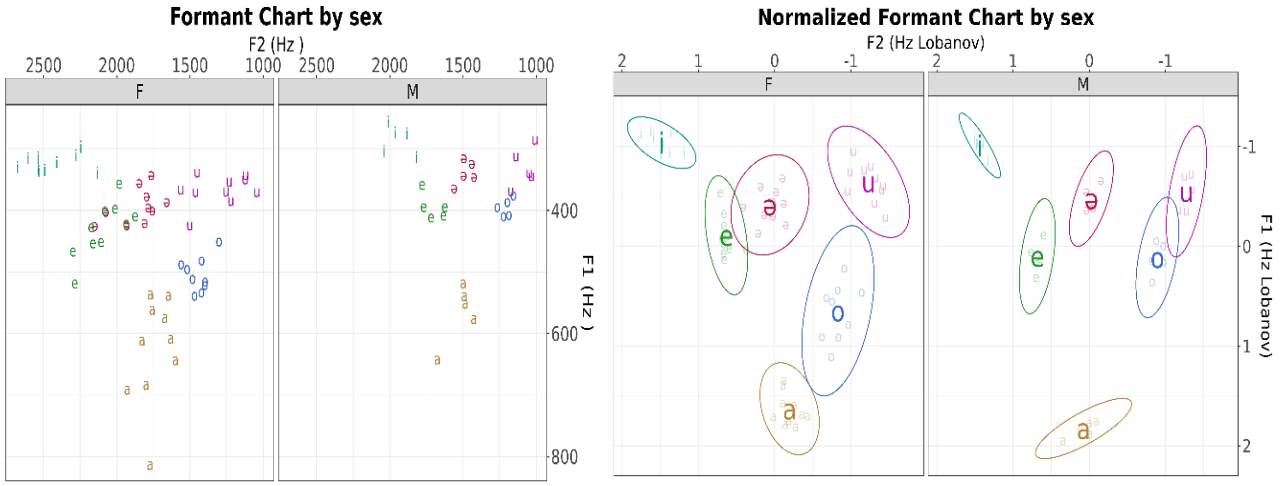



*Figure 1: Répartition des voyelles et du vocoïde sur l'espace acoustique : à gauche valeurs non-normalisées, à droite valeurs normalisées.*

### 3.2 Réalisations phonétiques de [ɾ]

Nous allons maintenant présenter les différentes réalisations du tap lors de la production du cluster étudié. Le tableau I ci-dessous, indique les différentes réalisations du cluster que nous avons rencontré lors de l'analyse de nos données ainsi que le nombre d'occurrences par réalisation.

*Tableau I : Synthèse des réalisations phonétiques, leur description et nombre d'occurrences.*

| Forme phonologique | Réalisations phonétiques | n=481 | Description Phonétique |
|---|---|---|---|
| /t͡ɾ/ | [t̪͡ɾ] | 224 | Occlusive dento-alvéolaire sourde + tap apico-alvéolaire occlusif voisé |
| | [t̪͡ɾ̥] | 35 | Occlusive dento-alvéolaire sourde + tap apico-alvéolaire occlusif sourd |
| | [t̪͡ɾ̞] | 128 | Occlusive dento-alvéolaire sourde + tap apico-alvéolaire approximant bref |
| | [t̪͡ɹ] | 50 | Occlusive dento-alvéolaire sourde + tap apico-alvéolaire approximant long |
| | [t̪͡s̪] | 44 | Affriquée lamino-alvéolaire sourde |
| | [s̪] | 1 | Fricative lamino-alvéolaire sourde |

Nous avons pu observer six réalisations différentes de la rhotique parmi lesquelles paraît prédominer la réalisation canonique du tap occlusif [ɾ] (46%), suivi de la réalisation approximante brève [ɾ̞] (26%). On notera que, dans 2% des cas, la rhotique peut s'élider ou se réaliser comme [ɚ̯] (notons qu'il ne s'agit aucunement de la rhotique anglaise mais d'une transcription soulignant l'assimilation entre le vocoïde et le tap) dont nous ne parlerons pas ici.

Les taps [ɾ ɾ̥] se caractérisent acoustiquement par une brève occlusion apico-alvéolaire se traduisant par un silence sur le spectrogramme, et parfois (pas toujours) par un relâchement visible sur le spectrogramme, sous la forme d'un bruit d'explosion. Cependant, ils diffèrent en deux points : leur durée et la présence ou l'absence d'une barre de voisement dans les basses fréquences du spectrogramme et des oscillations simples sur le signal sonore. Le tap voisé [ɾ] a une durée moyenne de 22,5ms chez les locuteurs AB mais de 38ms chez les locutrices C

(hyperarticulation) et son corrélat sourd [ɾ̥] a une durée de 28,5ms chez les locuteurs AB mais de 37ms chez les locutrices C (encore une fois il y a hyperarticulation).

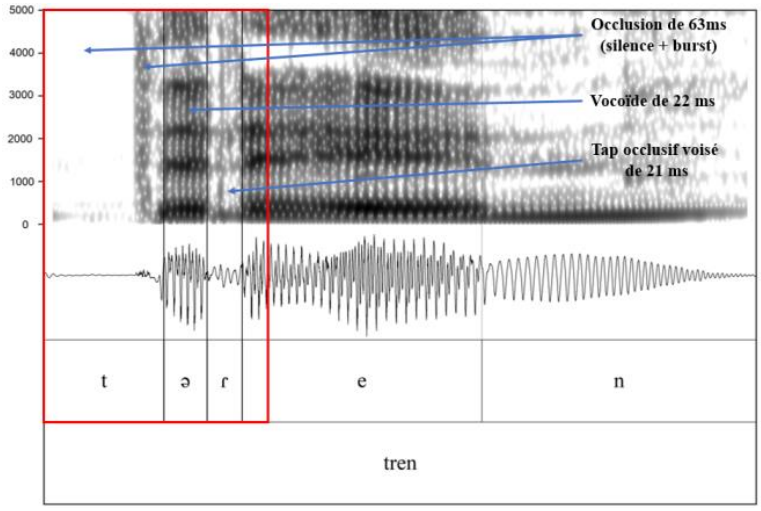



*Figure 2: Spectrogramme du mot "tren" (train) par une locutrice de 28 ans (échelle = 5000Hz)*

Les taps approximants [ɾ̬ ɹ] se différencient des taps [ɾ ɾ̥] par le degré de contact des articulateurs en jeu. Dans le cas des approximantes, les articulateurs ne se touchent pas et sont trop éloignés pour permettre la production d'une turbulence ; ils se caractérisent donc par la présence de formants d'intensité plus faible que pour des voyelles. Les deux taps [ɾ̬ ɹ] se distinguent par leur durée : 22ms pour l'approximante brève [ɾ̬] tous locuteurs confondus (pas d'hyperarticulation cette fois-ci) et 35ms (locuteurs AB) contre 41ms (locutrices C, clairement hyperarticulé) pour l'approximante longue [ɹ].

Enfin, nous avons pu constater l'existence d'une réalisation fricative de la rhotique que l'on note [ɻ̊]. Cette réalisation renvoie à deux allophones, d'une part [t͡ɻ̊] (affriquée lamino-alvéolaire sourde selon la description de Sadowsky, 2015) et [ɻ̊] d'autre part (qui est en fait le résultat de la perte du caractère occlusif de l'affriquée). Pour ce qui de [t͡ɻ̊], la littérature (Sadowsky, 2015 ; Figueroa *et al.*, 2013) fait la distinction entre trois réalisations de cette affriquée en fonction du rapport entre la durée de l'occlusion/friction (e.g. [t͡ɻ̊] [t͡ɻ̊] [t͡ɻ̊]) et la stigmatisation sociolinguistique qui en est faite. Dans ce travail nous avons laissé de côté ce détail car nous ne travaillons pas sur les préjugés sociolinguistiques et attitudes négatives qui en découlent. La durée moyenne de de la rhotique fricative de l'allophone affriqué va de 9ms jusqu'à 65ms et dure en moyenne 46ms. Sur les 44 occurrences rencontrées, 61,4% des affriquées sont produites par les locutrices C et 38,6% par les locuteurs AB. On retrouve l'allophone [t͡ɻ̊] dans différentes contextes segmentaux et prosodiques sans qu'aucun pattern homogène et linguistiquement explicable ne s'esquisse. Une seule occurrence de l'allophone [ɻ̊] a été rencontrée dans le mot « atrasado » 'en retard' ([aɻ̊aˈsao]) produit par une locutrice de 61 ans du groupe AB. L'analyse spectrographique n'indique aucune occlusion avant la rhotique fricative ; le segment a une durée de 76 ms.

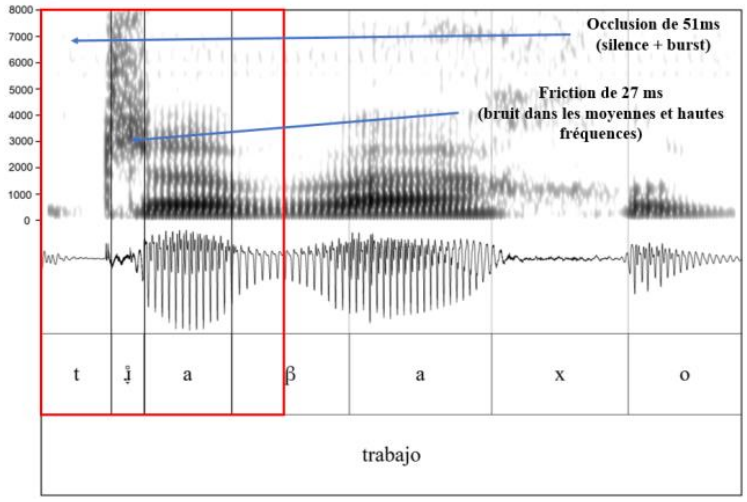



*Figure 3 : Spectrogramme du mot "trabajo" (travail) par une locutrice du groupe C (ɻ̊ désigne la rhotique fricative [ɻ] sourde [˚])*

## 4. Discussion

Les données de la distribution des voyelles sur l'espace acoustique montrent un système vocalique centralisé comme le suggère Sadowsky (2012). Cette réduction de l'espace acoustique, par rapport à d'autres variantes de l'espagnol (Sadowsky, 2012), pourrait s'expliquer par l'influence des voyelles du mapudungun dont le système est centralisé (*ibid.*).

Nous avons vu que le vocoïde s'insérait entre les deux consonnes coronales (dans 86% des cas) ; dans notre interprétation des données, nous rejoignons celle de Fougeron & Ridouane (2008) : l'élément vocalique est un vocoïde de transition qui joue un rôle dans la coordination des gestes articulatoires de structures de type CC, dans notre cas cela permet de réduire le chevauchement entre les gestes et ainsi éviter les effets d'une coarticulation entre /t̪/ et /ɾ/ (i.e. empêcher l'assimilation des deux segments en [t̪̊ɻ̊]). Le vocoïde n'est pas épenthétique mais intrusif car il ne peut assumer la fonction de noyau de syllabe (Hall, 2006). La syllabation de « trampa » 'piège' donnera /t̪ɾam.pa/ → [t̪ᵊɾam.pa] et non *[t̪ᵊ.ɾam.pa]. Cet argument prouve aussi que le vocoïde n'est pas défini au niveau phonologique mais plutôt lors de l'implémentation phonétique.

À la lumière des données obtenues, nous confirmons ce qui apparaît dans la littérature : l'affrication et la spirantisation du cluster /t̪ɾ/ n'est pas entièrement motivée par des facteurs linguistiques mais par des facteurs extralinguistiques. Deux facteurs sociolinguistiques nous ont interpellé : l'âge et le profil socio-économique (la variable de sexe n'a pas été traitée car nous n'avons pas de locuteurs masculins du groupe C pour établir une comparaison viable). Sur 44 occurrences de [t̪̊ɻ̊], 61.4% ont été produites par des locutrices du groupe C ; pour le même nombre d'occurrences, 68% ont été produites par des locutrices de plus de 50 ans. Les approches phonologiques traditionnelles n'incluent pas la variation extralinguistique dans leurs théories. Sebregts (2015) se propose d'étudier la variation sociophonétique et phonologique du « *Dutch –r* » en s'appuyant sur des théories qui intègrent la variation : la théorie à exemplaire, l'approche diachronique, la phonologie articulatoire. Il n'existe pas, à notre connaissance, d'étude diachronique sur /t̪ɾ/ ; on sait cependant qu'à la fin du 19ᵉ siècle [t̪̊ɻ̊] était principalement produit par des locuteurs analphabètes des classes populaires (Lenz, 1940 [1892–93]). La théorie à exemplaire se fonde non pas sur des catégories phonémiques mais sur des catégories lexicales ; les représentations lexicales se forment à partir des différents exemplaires perçus d'un même mot. Ainsi, la variation, les détails phonétiques fins et les facteurs extralinguistiques (e.g. âge, profil socio-économique, sexe, dialecte etc…) sont stockés en mémoire. La structuration phonologique

découle de la généralisation faite des exemplaires. La production se fait par sélection d'un exemplaire gardé en mémoire. On peut tout à fait penser que les formes [t̪ʰɾ] et [t͡s̪] (parmi les allophones du groupe) constituent les attaques complexes des exemplaires : [t̪ʰɾ] serait la réalisation standard (c'est-à-dire défini comme la norme) et [t͡s̪] la réalisation dominante dans le sociolecte des locutrices du groupe C. La phonologie articulatoire (Browman, C. & Goldstein, L., 1990) va nous permettre de comprendre l'organisation spatiotemporelle des gestes de [t̪ʰɾ] et [t͡s̪] une fois l'exemplaire choisi. Pour la réalisation de [t͡s̪], un chevauchement maximal (Bradley, 2002 p.7) va être planifié afin que les consonnes, partageant (pratiquement) la même zone articulatoire, s'assimilent pour donner un phone affriqué ; à l'inverse, dans le cas de [t̪ʰɾ], un chevauchement partiel va avoir lieu permettant ainsi l'intrusion d'un vocoïde et qui va en empêcher l'assimilation consonantique.

Les données recueillies montrent que les locutrices du groupe C allongent la durée de certains segments, cela nous a invité à nous interroger sur ces données : hyperarticulation ou hypercorrection ? Lorsque quelqu'un hyperarticule, il accentue les traits phonétiques d'un mot, quand un instituteur fait une dictée par exemple. L'hypercorrection, elle, résulte de la production exagérée d'une forme considérée comme plus prestigieuse ou plus proche de la norme. Lorsque l'on hypercorrige sa production on bascule du linguistique au sociolinguistique. Les études de Figueroa et Ñañucleo (2013) et Sadowsky (2015) soulignent que les locuteurs chiliens adultes, plus âgés et de bas profil socio-économique utilisent principalement les allophones affriqués [t͡s̪] [t͡ɬ̪] [ɬ̪], lesquels sont fortement stigmatisés par les strates moyennes et hautes. D'autres recherches ultérieures permettraient d'éclairer ces questions, ici nous ne pouvons y répondre ; pour cela il serait utile de procéder à un test de perception pour évaluer le jugement porté à ces variantes allophoniques.

## 5. Conclusion

Le tap apico-alvéolaire est un segment complexe qui se définit phonologiquement comme une occlusive de très courte durée (i.e. 22 ms), c'est-à-dire /ɾ/, qui devrait se traduire acoustiquement par une occlusion, un silence, une barre de voisement et un relâchement visibles sur le signal et su le spectrogramme. On le retrouve généralement à l'intervocalique ce qui souligne la nécessité pour ce phone de se trouver dans un environnement vocalique d'où la nécessité du vocoïde dans les formes de type [t̪ʰɾV], soulignant bien entendu la relation indissociable de ces deux éléments dans le cluster. Au niveau phonétique, sa réalisation est plus complexe étant donné que /ɾ/ répond à un ensemble de facteurs articulatoires et perceptuels afin que cet élément gagne en saillance. C'est un phone versatile et complexe qui tantôt est vocalique (e.g. [ɔ̯ ɾ̥ ɹ]) tantôt consonantique (e.g. [ɾ ɾ̥ ɬ̪̆]). Nos données montrent une prédominance du tap (occlusif ou approximant), de plus l'hypercorrection (expliquée ci-dessus) de [t͡s̪] en [t̪ʰɾ] suggère un usage normatif de la forme [t̪ʰɾ] qui, du point de vue phonologique, doit être spécifiée comme /t̪ɾ/. En attaque complexe on voit donc apparaître un continuum de complexité articulatoire motivé par des facteurs sociolinguistiques :

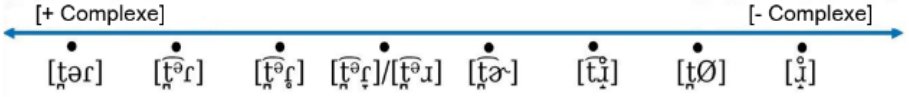



*Figure 4: échelle de complexité articulatoire.*

Ce phénomène n'est pas propre à l'espagnol du Chili étant donné que l'on retrouve des sons similaires dans d'autres variétés d'espagnol mais aussi dans d'autres langues : [t͡ɬ̪] en anglais, en Sicilien ou [t͡ʂ] en Mandarin, Mapudungun. Il est tout à fait concevable que cela soit le résultat

d'un mécanisme universel où une rhotique (qui se trouve dans la région prépalatale ou alvéolaire) au contact d'une occlusive dentale/alvéolaire se spirantise pour des questions physiologiques et de VOT.

**Remerciements**

L'auteur remercie chaleureusement Didier DEMOLIN pour son aide tout au long du travail ainsi qu'Élodie BLESTEL pour ses remarques et corrections et tous ses locuteurs.

## Références

ALONSO, A. (1953). *"La pronunciación de "rr" y de "tr" en España y América": Estudios lingüístico. Temas hispanoamericanos.* Madrid: Gredos. 151-195.

ASOCIACIÓN CHILENA DE EMPRESAS DE INVESTIGACIÓN DE MERCADO [AIM]. (2008). *Descripción de grupos socioeconómicos.* Santiago. En ligne sur http://www.aimchile.cl/wp-content/uploads/2011/12/Grupos_Socioeconomicos_AIM-2008.pdf.

BRADLEY, T. G. (2002). Gestural Timing and the Resolution of /Cr/ Clusters in Romance. *Linguistic Symposium on Romance Languages (Vol. 32)*, 19-21.

BRADLEY, T. G. (2004). Consonantes róticas : descripción fonética. Dans J. G. (eds.), *Fonética y fonología descriptiva de la lengue española* (p. Chap. 21). Madrid: Consejo Superior de Investigación Scientífica.

BROWMAN, C. & GOLDSTEIN, L. (1990). Tiers in Articulatory Phonology, with Some Implications for casual speech. Dans B. &. (eds.), *Papers in Laboratory Phonology I: between the grammar and the physics of speech.* (pp. 341-376). Cambridge, U. K.: Cambridge University Press.

CICRES, J. ET BLECUA, B. (2015). Caracterización acústica de las róticas fricativas prepausales en español peninsular. *Loquens 2 (1)*, 1-12.

FIGUEROA, M., SALAMANCA, G., ET ÑANCULEO, M. (2013). El eje oclusión-fricción en el sistema sociofónico del castellano chileno. *Estudios de Fonética Experimental*, 233-273.

FIGUEROA, M., SOTO-BARBA, J. Y ÑANCULEO, M. (2010). Los alófonos del grupo consonántico /tr/ en el castellano de Chile. *Onomázein 22*, 11-42.

FOUGERON, C., & RIDOUANE, R. (2008). On the nature of schwa-like vocalic elements within some Berber clusters. *Proceedings of the eighth international seminar on speech production*, (pp. 441-444).

HALL, N. (2006). Cross-linguistic patterns of vowel intrusion. *Phonology, 23(3)*, 387 - 429.

LENZ, R. (1940 [1892–93]). El español en Chile. Dans A. A. (eds), *El español en Chile. Trabajos de Rodolfo Lenz, Andrés Bello y Rodolfo Oroz* (pp. 80–268). Buenos Aires: Facultad de Filosofia y Letras de la Universidad de Buenos Aires.

MELLA, B. (2009, Août 29). *Dime dónde vives y te diré quien eres, una radiografía a la sociedad santiaguina.* Récupéré sur Plataforma Urbana: http://www.plataformaurbana.cl/archive/2009/08/29/dime-donde-vives-y-te-dire-quien-eres-una-radiografia-a-la-sociedad-santiaguina/

SADOWSKY, S. (2012). Vocales de referencia del castellano de Chile. *V Jornadas Nacionales de Fonética : Temuco, Chile*.

SADOWSKY, S. (2015). Variación sociofonética de las consonantes del castellano chileno. *Sociolinguistic Studies 9 (1)*, 71-92. doi:10.1558/sols.v9i1.19927

SEBREGTS, K. (2015). *The Sociophonetics and Phonology of Dutch r (PhD dissertation).* University Utrecht.

# Étude exploratoire des stratégies de production du ton 3 en chinois mandarin

Yizhi Huang[1], Véronique Delvaux[1,2,] Kathy Huet[1], Myriam Piccaluga[1], Guoxian Zhang[1], Bernard Harmegnies[1]

(1) Institut de Recherche en Sciences et Technologies du Langage, UMONS, Belgique
(2) FNRS, Belgique
yizhi.huang@umons.ac.be

## RESUME

Malgré l'absence de contraste linguistique au niveau du mode de phonation en chinois mandarin, des études antérieures ont documenté, lors de la production des tons, de fréquentes occurrences de phonation non modale (type "creaky voice"/"vocal fry"), en particulier avec le troisième et le quatrième ton. La f0 étant traditionnellement considérée comme l'élément déterminant dans la production des tons, leurs autres corrélats acoustiques/phonétiques demeurent inexploités dans ces études antérieures. Sur la base d'un corpus constitué de mots et de pseudo-mots monosyllabiques de structure CV (où V est [a] (tons T1, T2, T3, T4), et C [m]) produits isolément par 10 locuteurs natifs du chinois mandarin, la présente étude vise à décrire la distribution des occurrences de phonation non modale en fonction des paramètres structurant le corpus, et à mieux caractériser les stratégies de production du T3 en particulier, à partir de mesures acoustiques diversifiées.

## ABSTRACT

**Preliminary study on strategies of native T3 production in Mandarin Chinese**
Non-modal phonation is present in Mandarin Chinese speech production but marks no linguistic contrast. In previous studies, T3&T4 are reported to sometimes be produced with creaky voice/vocal fry. The pitch is normally considered as the primary property in production, but other acoustic/phonetic correlates of the Mandarin Chinese lexical tones remain not fully understood. Basing on a corpus consisted of monosyllabic words and non-words produced in citation form by 10 native speakers of Mandarin Chinese, the present study aims at describing the occurences of non-modal phonation in Mandarin Chinese, particularly in T3 production, and the characteristics of the strategies which are implemented in tone production with various acoustic measurements.

MOTS-CLES : tons; chinois mandarin; voix craquée; phonation
KEYWORDS : tones; mandarin; creaky voice; phonation

# 1 Introduction

Le chinois mandarin est doté d'un système de contraste tonal à cinq tons, dont un ton statique (T1), trois tons dynamiques (T2, T3 T4) et un ton neutre. Une méthode de transcription très adoptée par


10.21437/JEP.2018-62

les chercheurs du domaine qu'est la numérotation de *Chao* (Chao, 1930) référencée comme Chao digit par certains phonologues (Yip.M., 2003 ;Duanmu. S, 2007, 226) où le registre et le contour de pitch sont représentés sur une échelle à cinq niveaux numérotés de 1 à 5, du registre le plus bas vers le plus haut. Les quatre tons contrastifs du mandarin peuvent ainsi être transcrits comme 55 (T1), 35 (T2), 214 (T3), 51 (T4), ce qui permet de marquer le registre et la dynamique de pitch d'un ton. Traditionnellement, la fréquence fondamentale (f0 moyenne, évolution temporelle de la f0) est considérée comme le principal corrélat acoustique des tons. Quelques études ont évoqué d'autres indices phonétiques/acoustiques qui pourraient jouer un rôle dans la production et la perception des tons - parmi lesquels le voisement de la consonne précédente, la durée syllabique et la qualité vocale/le mode phonatoire - mais ces éléments demeurent globalement peu exploités dans la littérature décrivant la production et, plus encore, la perception des tons (Kuang, 2017).

En ce qui concerne le rapport entre mode phonatoire et tons, des études antérieures ont suggéré une covariation entre la phonation non-modale de type *creaky* et le pitch en chinois mandarin, le T3 et le T4 étant parfois réalisés avec une qualité de voix dite *« creaky voice »* (Chao, 1956 ; Davidson, 1991; Belotel-Grenié et Grenié, 1994; Kuang, 2017). Rappelons qu'en chinois mandarin, la phonation non-modale de type « creaky voice » ne supporte pas de contraste linguistique (Moisik et al., 2014). Il s'agirait plutôt ici de variation allophonique typiquement associée à une fréquence fondamentale très basse (Kuang, 2017). Le chinois mandarin n'est pas la seule langue présentant ce type de covariation. En vietnamien du nord (Brunelle, 2009) et en cantonais (Yu et Lam, 2014), ce mode phonatoire particulier est employé de manière allophonique avec les tons. Yu et Lam (2014) ont étudié le rôle du « creaky voice » dans la perception des tons en cantonais, et ont montré une amélioration de 20% du taux d'identification du T4 (un ton avec un registre bas) lorsque les stimuli étaient accompagnés du mode phonatoire « creaky » par rapport aux stimuli produits en phonation modale. Une étude récente (Yang, 2011) a également montré que l'emploi de « creaky voice » est particulièrement utile pour discriminer le T2(35) et T3(214), la paire tonale la plus confondue en perception et en production chez les natifs du mandarin, car le T3 (214) est souvent produit avec un profil tonal ressemblant à celui du T2(35) dans le contexte sandhi T3T3 ou T2T3. Puisque seul le contexte monosyllabique est présent, la particularité du ton neutre (il s'agit d'une absence de ton pour les syllabes non accentuées) n'est pas examinée ici.

L'objectif de la présente étude exploratoire est de décrire acoustiquement la production du T3 à partir des productions de 10 natifs du chinois mandarin, avec une attention particulière portée à la variabilité inter-individuelle et aux potentielles stratégies impliquant l'emploi d'indices acoustiques complémentaires à l'évolution de la fréquence fondamentale. Premièrement, l'évolution de la f0 sera étudiée et comparée à l'attente phonologique (T3=214), tant au point de vue du groupe que des profils individuels. Ensuite, on s'intéressera à la présence de très basses fréquences et à leur éventuel corollaire, le mode phonatoire de type "creaky voice". Enfin, on recherchera la présence de corrélats acoustiques autres que ceux associés à la f0 lors de la production du T3, principalement les variations de timbre et de durée vocalique. L'orientation de cette étude exploratoire est justifiée par notre interrogation sur l'importance relative des diverses stratégies de réalisation des tons du chinois mandarin, et ce dans une perspective ultérieure d'implémentation des observations réalisées auprès des locuteurs natifs dans le contexte de développements en didactique du Chinois langue étrangère à l'intention de natifs du français. Les données recueuillies dans la présente étude seront par ailleurs réutilisées dans de futures expériences en perception des tons lexicaux par des apprenants francophones.

# 2 Méthodologie

## 2.1 Corpus et participants

Le corpus est constitué de mots et de pseudo-mots monosyllabiques de structure CV où C est [m] et V est [a], portant les 4 tons lexicaux: T1, T2, T3, T4. Les items du corpus ont été répétés 4 fois de façon isolée par 10 natifs du chinois mandarin, 5 hommes et 5 femmes âgés de 18 à 36 ans, dont la plupart venant des régions où les dialectes de la famille mandarin (Pékin 1, Hebei 1, Henan 1, Shandong 2, Heilongjiang 1, Nanjing 2, Hubei 1) sont parlés, y compris le mandarin pékinois. Seule personne parlant un dialecte (Wu) hors de la famille des dialectes mandarin est l'auteur de l'article, ayant obtenu le certificat national de Putonghua. Les sessions d'enregistrement se sont déroulées dans une chambre sourde, les participants devant produire des stimuli en idéogrammes et pinyin affichées sur un écran d'ordinateur en face d'eux. La prononciation et le sens des idéogrammes moins connu sont préalablement expliqués à chaque participant. Les enregistrements ont été effectués à l'aide d'un enregistreur digital zoom H5.

## 2.2 Mesures

Les mesures ont été effectuées manuellement à l'aide du logiciel Praat. La fréquence fondamentale (f0, en Hz) et les 4 premiers formants (F1, F2, F3, F4, en Hz) ont été mesurés en début et fin des syllabes portant le ton 3, (les données recueillies sur les trois autres tons sont utilisées uniquement en comparaison avec la durée moyenne du T3 dans le présent article). La durée de la syllabe ainsi que la durée de la consonne initiale ont également été mesurées. Enfin, on a codé chaque occurrence de T3 pour la présence (1) ou l'absence (0) de "creaky voice" sur base de critères auditifs (jugement expert par le premier auteur, natif du mandarin) et visuels: présence d'irrégularités au niveau des pulses glottaux, détection d'un pitch discontinu par Praat, changement important d'amplitude et/ou de fréquence entre cycles consécutifs visibles sur le signal de parole et/ou le spectrogramme. Notons qu'en cas de rupture dans la détection de f0 par l'algorithme de pitch de Praat, bien que la périodicité des pulses glottiques demeure, la f0 a été évaluée directement à partir du signal de parole sur la période la plus longue, via la formule: $f0(Hz)=1/T(ms)$.

# 3 Résultats

## 3.1 Fréquences fondamentales moyennes

Comme le montre la figure 1, les fréquences fondamentales du T3(214) observées sont, en moyenne, plus élevées chez les femmes que chez les hommes, comme on pouvait évidemment s'y attendre, étant donné les différences de conformation physique. On observe, de manière plus intéressante, que les valeurs calculées présentent des tendances cohérentes avec l'attente phonologique puisque,

tant chez les locuteurs féminins que chez les locuteurs masculins, apparaît la forme en V, qui trahit une baisse sensible du pitch au centre de la syllabe.

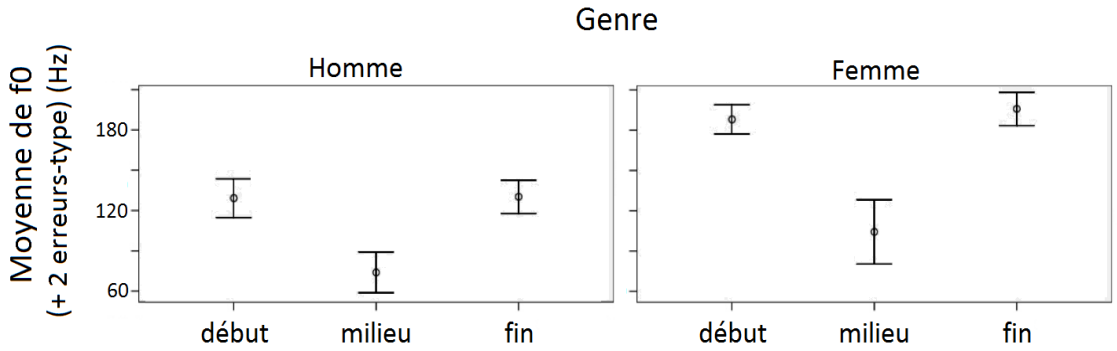



*Figure 1 : Moyenne et intervalle de confiance (moyenne ± 2 erreurs types) des valeurs de fréquence fondamentale (en Hz) relevées en début, milieu et fin de syllabe portant le T3 pour les hommes (en haut) et les femmes (en bas).*

Dans chacun des deux groupes de genre, le fondamental moyen de fin de syllabe est, par ailleurs, supérieur à celui du début (hommes : 130,20 Hz > 129,25 Hz; femmes : 195,85 Hz > 188,05 Hz) ; cette différence est cependant extrêmement ténue.

En outre, si les valeurs de début et de fin de syllabe sont certes légèrement élevées pour chacun des groupes de genre mais cependant assez banales (de l'ordre de 130 Hz chez les hommes et de 190 Hz chez les femmes), elles sont spécialement basses en milieu de syllabe (74 Hz dans le groupe masculin et 104 Hz dans le groupe féminin). On constate aussi que les variances sont maximales au centre de la syllabe (femmes : $\sigma$= 53 Hz > 24 Hz et 27 Hz; hommes : $\sigma$= 33 Hz > 32 Hz et 27 Hz).

## 3.2 Fréquences fondamentales: variabilité interindividuelle

La figure 2 permet un regard plus précis sur la variabilité interindividuelle caractérisant les valeurs de fréquence fondamentale. Dans chaque groupe de genre, elle présente, pour chaque sujet, ses valeurs de f0 mesurées en début et fin de syllabe (reliées par un segment de droite), et sa valeur mesurée en milieu de syllabe (représentée par un cercle), le tout rassemblé en un rectangle vertical grisé. La figure 2 permet de prendre la pleine mesure de l'importante variabilité interindividuelle présente tant dans le groupe masculin que dans le groupe féminin, les sujets se dispersant largement sur l'axe des fréquences fondamentales.

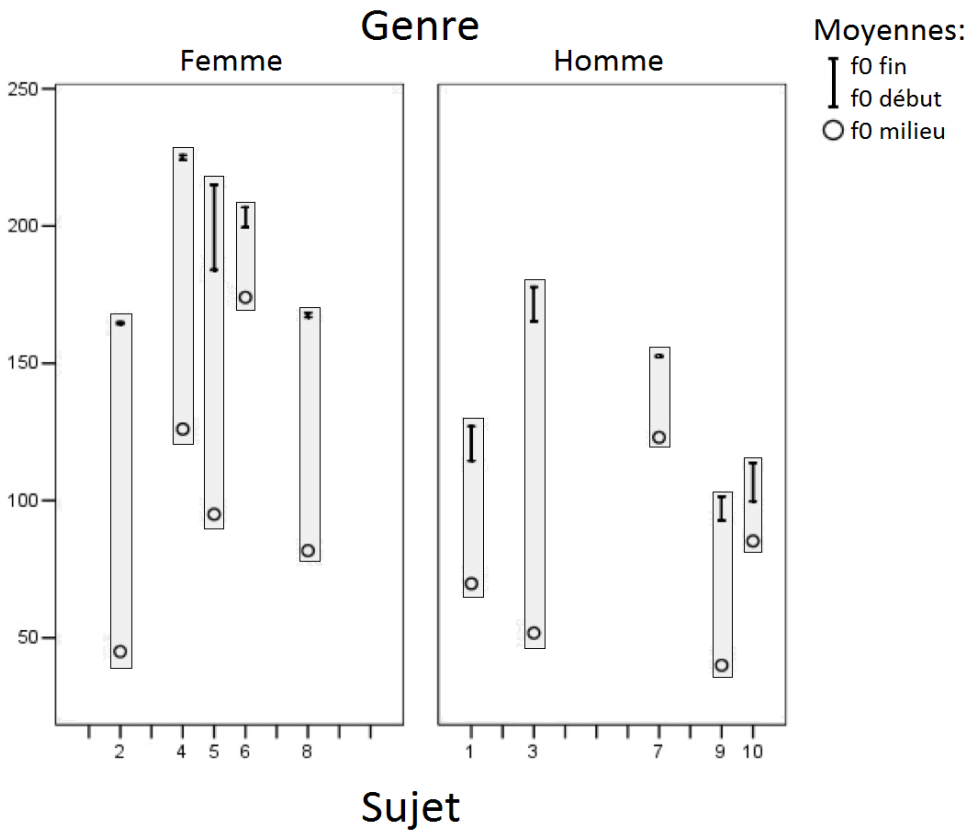



*Figure 2 : Fréquence fondamentale (en Hz) en début et fin de syllabe (valeurs reliées par un segment de droite) ainsi qu'en milieu de syllabe (valeur figurée par un cercle), présentées verticalement pour chacun des sujets des deux genres (moyennes sur 4 productions).*

Par-delà ces différences interpersonnelles, la figure fait cependant apparaître des récurrences. On observe ainsi que souvent, le segment se réduit pratiquement à un point, les valeurs de début et fin de syllabe étant similaires sinon égales. Là où apparaît un segment de droite (c'est à dire en cas d'inégalité des valeurs initiale et finale), il est pratiquement toujours de faible ampleur en comparaison de la distance entre son point le plus bas et le cercle figurant la valeur de milieu de syllabe. Dans tous les cas, le cercle apparaît en-dessous du segment (ou du point), ce qui confirme, pour les individus considérés isolément, le constat global effectué plus haut (Figure 1) d'une infériorité du pitch de milieu de syllabe par rapport aux valeurs qui l'entourent.

Quelque intéressants que soient ces constats, ils ne permettent cependant pas d'accéder aux caractéristiques des réalisations isolément, puisque les données présentées dans la figure 2 sont des moyennes calculées sur 4 productions.

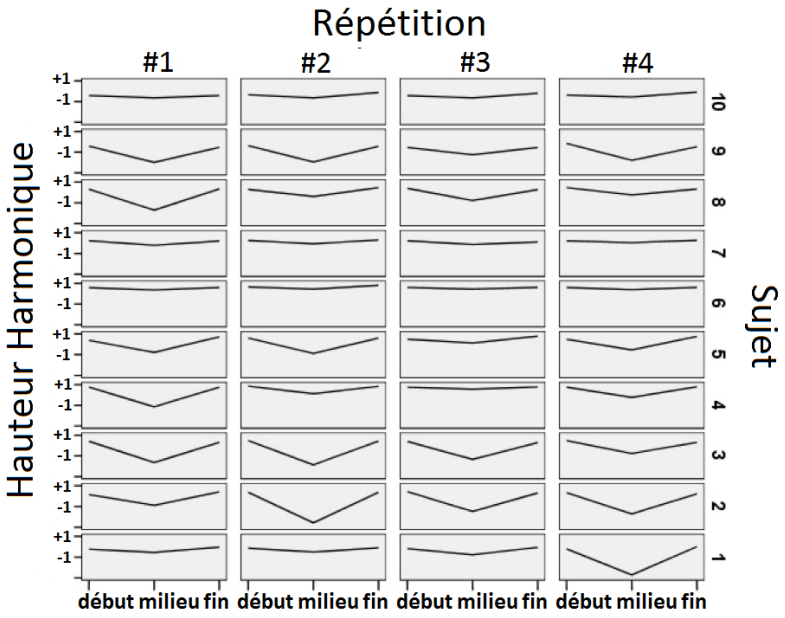



*Figure 3 : évolution de la mélodie pour chaque production de syllabe, et chaque sujet (pitch en Hauteurs harmoniques).*

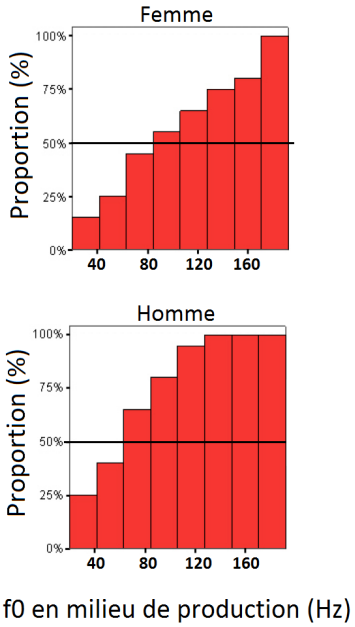



*Figure 4 : distributions cumulées des valeurs centrales de f0 (Hz): femmes et hommes.*

Dans la figure 3, par contre, les profils d'évolution de la fréquence fondamentale dans la syllabe sont détaillés sujet par sujet et production par production. Afin de rendre plus comparables les profils tonaux (vu l'importante variabilité interindividuelle observée), les fréquences en Hertz (F) ont ici été converties en hauteurs harmoniques (Ha) selon la formule (Pierart B., Harmegnies B., 1993) :

$$Ha = (1/LOG10\ (2)) * LOG10\ (F/131)$$

Des profils de sujets très différents apparaissent ici. On observe, par exemple, que les sujets 2, 3, 8 et 9 opèrent systématiquement une diminution sensible du pitch en milieu de syllabe. A contrario, les sujets 6, 7 et 10 présentent des profils d'évolution pratiquement plats. Certains sujets, quant à eux, se montrent inconstants de production à production, affichant tantôt un profil plat, tantôt un profil en V.

### 3.3 Occurrences de (très) basses fréquences

Les données présentées plus haut indiquent la présence de fréquences très basses tant dans le groupe masculin que dans le groupe féminin. A ce propos, la figure 4 montre des histogrammes cumulés qui, construits au départ des valeurs individuelles de fréquences fondamentales de milieu de syllabe, rendent compte des valeurs de f0 avec la plus fine des granularités possible. On peut y observer un nombre de très basses fréquences (même inférieures à 50 Hz) important et ce, tant chez les hommes que chez les femmes. La figure 5 illustre le phénomène en présentant le sonagramme d'une séquence /ma/ produite avec un ton T3 par un jeune locuteur La fréquence fondamentale est de

l'ordre de 170 Hz tant en début qu'en fin de production, mais elle chute considérablement en milieu de production. Les pulses glottiques s'espacent et se raréfient graduellement, puis se resserrent tout aussi progressivement. Au centre, l'écart interpulses constaté est de 20 ms, ce qui correspond à une fréquence de 50Hz. Sur le plan perceptif, la zone centrale est systématiquement interprétée comme un épisode de voix craquée.

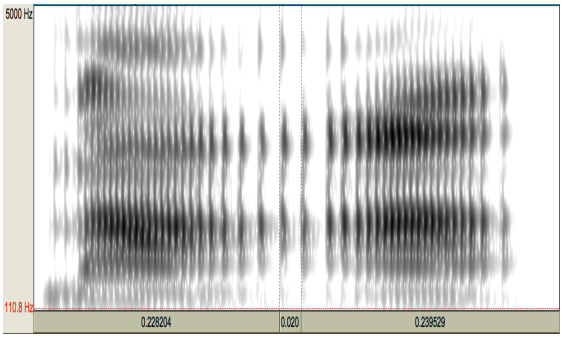



*Figure 5 : sonagramme d'une réalisation de la syllabe /ma/ (T3), avec f0 très basse (≅50 Hz).*

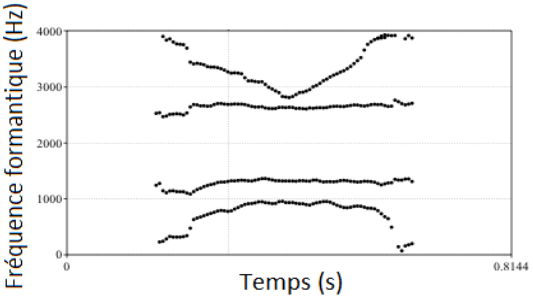



*Figure 6 : graphe formantique de la réalisation de la syllabe /ma/ (T3, sujet 7, production #2), avec éclaircissement du timbre.*

## 3.4 Autres corrélats acoustiques du T3

L'examen des valeurs des formants indique que certains sujets recourent (occasionnellement ou systématiquement) à un éclaircissement du timbre lors de la production du T3. Celui-ci est illustré dans le graphe formantique (Cf. Figure 6) par la production 2 du sujet 7 (celui de notre échantillon qui présente de la manière la plus nette et la plus systématique ce profil). Comme on peut le voir, c'est la dynamique du quatrième formant qui semble ici déterminante, sa fréquence passant de 2800 Hz en milieu de syllabe à 4100 Hz en fin de syllabe. Dans une étude récente (Zhang, 2017) sur les tons lexicaux du Chinois Mandarin en voix modale et en voix chuchotée, on observe que la dynamique du quatrième formant est parfois plus net, en liaison avec la nature du segment vocalique et le mode de phonation, que celui des trois premiers formant. Trois des sujets de notre échantillon présentent cette tendance (sujets 3, 7 et 10). Pour deux d'entre eux, les réalisations ne comportent en effet pratiquement pas de variation mélodique, ce qui suggère une stratégie compensatoire fondée sur la variation du timbre pour produire la dynamique nécessaire à l'implémentation du T3, qui repose ici plus sur l'évolution du contenu spectral que sur celle de la fréquence fondamentale.

Enfin, les mesures de durée syllabique effectuées sur l'ensemble du corpus (T1, T2, T3, T4) font apparaître des différences significatives de durée en fonction du ton, confirmées par l'analyse de variance ($F=4625{,}797$, $df=3$, $p<.001$). Pour chacun des 10 sujets, le ton 3 est systématiquement porté par une syllabe plus longue. Cette différence de durée est cependant plus marquée chez certains locuteurs que chez d'autres, ce qui pourrait suggérer l'existence de stratégies individuelles de production du T3 recourant à ce paramètre.

# 4 Discussion et Conclusion

Dans cette étude exploratoire, nous avons décrit acoustiquement la production du T3 au départ de productions isolées de mots et pseudo-mots monosyllabiques par 10 locuteurs natifs du chinois mandarin. En ce qui concerne l'évolution temporelle de la fréquence fondamentale, on retrouve globalement la forme de "V" à laquelle on pouvait s'attendre étant donné la transcription phonologique traditionnellement associée à ce ton (214). Notre étude aboutit cependant à nuancer ce constat sur deux points. Premièrement, les valeurs de fréquence fondamentale en début et fin de T3 divergent peu l'une de l'autre et ne justifient certainement pas un "saut" de deux registres (de 2 à 4). Deuxièmement, la forme en "V" résume adéquatement les données considérées dans leur ensemble mais ne caractérise pas chaque production réalisée par chaque locuteur.

Se pose dès lors la question des autres corrélats acoustiques potentiels de la réalisation du T3, outre l'évolution temporelle de la f0. Dans cette étude, nous en avons considéré trois. Tout d'abord, en accord avec la littérature récente (Kuang, 2017), nous avons régulièrement constaté la présence de fréquences extrêmement basses dans la phase médiane de la réalisation du ton, qui étaient associées sur le plan perceptif à un épisode de voix craquée. Il faut souligner, cependant, que nous n'avons observé ici aucune interruption de la périodicité, ni aucun bruit venant se surajouter au phénomène, ce qui caractérise généralement le type "creaky voice" (p.ex. Keating et al., 2015). La question du mécanisme phonatoire sous-jacent demeure donc à ce stade. Il ne semble pas possible en tout cas qu'une si faible fréquence fondamentale puisse être obtenue à l'aide des seuls plis vocaux. L'hypothèse qu'une structure plus massive est ici mobilisée, peut être en sus des plis (bandes ventriculaires ?) devrait être testée au moyen d'instrumentation articulatoire. On pourrait alors considérer la possibilité qu'un mode phonatoire particulier a été adopté par des membres d'une communauté linguistique donnée (ici, celle des mandarinophones) et que certains locuteurs ont acquis la capacité d'en faire usage avec plus ou moins de succès, mais loins d'être systématique, ce qui confirme l'étude de Moisik, Lin & Esling (2014).

En effet, un autre aspect des données analysées ici est la grande variabilité inter-individuelle observée dans l'utilisation d'indices acoustiques complémentaires susceptibles de supporter la réalisation phonétique du T3. Certains locuteurs présentent des épisodes de voix craquée, d'autres recourent à un éclaircissement du timbre, d'autres encore à une augmentation plus importante de la durée syllabique (ce qui n'est pas toujours le cas dans la production de la parole connectée, où la durée vocalique des tons est réduite quasiment tous à la même durée, le profil des tons modifié due à la transition aux frontières tonales. Cela crée donc la confusion entre certains paires de tons tels que la paire T2/T3), ce qui suggère une diversité de stratégies individuelles (compensatoires?) permettant de supporter phonétiquement le contraste phonologique entre le T3 et les autres tons du mandarin. Il importe de souligner ici que rien n'indique une quelconque exclusivité mutuelle des démarches stratégiques observées. D'une part, certains locuteurs paraissent tantôt recourir à une stratégie et tantôt pas (par exemple le sujet 1, qui ne recourt massivement à la stratégie mélodique que dans sa production 4). D'autre part, rien n'exclut que plusieurs stratégies puissent simultanément être utilisées par certains locuteurs dans certaines productions. Bien entendu, la question de l'efficacité de ces stratégies et / ou de leurs combinaisons, et celle d'une éventuelle hiérarchie entre les divers corrélats acoustiques ici considérés, devra être adressée au cours de futures études en perception de la parole.

# Références


BELOTEL-GRENIÉ, A., & GRENIÉ M. (1994). Phonation types analysis in standard Chinese. *Proceedings of Spoken Language Processing*, 343-346.

BRUNELLE, M. (2009). Tone perception in Northern and Southern Vietnamese. *Journal of Phonetics,* 37, 79–96.

CHAO, Y. R. (1930). A system of tone-letters. *Le Maître Phonétique,* 45, 24–27.

CHAO, Y. R. (1956). Tone, intonation, singsong, chanting, recitative, tonal composition and atonal composition in Chinese. *For Roman Jakobson: Essays on the Occasion of His Sixtieth Birthday*, édité par M. Halle, H. Lunt, H. McLean, & C. V. Schooneveld (Mouton Publishers, The Hague, The Netherlands), pp. 52–59.

DAVISON, D. S. (1991).An acoustic study of so-called creaky voice in Tianjin Mandarin. *UCLA Working Papers in Phonetics,* 78, 50-57.

DUANMU, S. (2007). *The phonology of standard Chinese*. Oxford University Press.

KEATING, P., GARELLEK, M., & KREIMAN, J. (2015). Acoustic properties of different kinds of creaky voice. *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, UK, pp. 0821.1–0821.5.

MOISIK, S.R., LIN, H., & ESLING, J.H. (2014). A study of laryngeal gestures in Mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (SLLUS). *Journal of the International Phonetic Association*, 44(1), 21-58.

KUANG, J. (2017). Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice. *The Journal of the Acoustical Society of America,* 142, 1693.

PIERART, B., & HARMEGNIES, B. (1993). Dysphasie simple de l'enfant et langage de la mère. *L'année psychologique*, *93*(2), 227-268.

YANG, R.X. (2011). The phonation factor in the categorical perception of Mandarin tones. *Proceedings of ICPhS XVII*, Hong Kong, China, 2204–2207.

Yip, M. (2002). *Tone*. Cambridge University Press.

YU, K.M., & LAM, H.W. (2014). The role of creaky voice in Cantonese tonal perception. *The Journal of the Acoustical Society of America*, 136(3), 1320–1333.

ZHANG, X.L. (2017). Les tons lexicaux du chinois mandarin en voix modale et en voix chuchotée. Thèse, 238.

XXXIIe Journées d'Études sur la Parole
4-8 juin 2018, Aix-en-Provence, France

# Impact de la détection de la parole pour différentes tâches de traitement automatique de la parole

Florent Desnous[1, 2] Anthony Larcher[1] Sylvain Meignier[1]
(1) LIUM, Le Mans Université, France
{florent.desnous, anthony.larcher, sylvain.meignier}@univ-lemans.fr
(2) Institute for Infocomm Research - A*STAR, Singapore

RÉSUMÉ

Dans cet article, nous proposons de comparer plusieurs systèmes de détection de la parole et leurs impacts sur deux tâches du traitement de la parole : la Segmentation et le Regroupement de Locuteurs (SRL) et la Reconnaissance Automatique de la Parole (RAP). Des systèmes à base de mixtures de Gaussiennes (GMM), de réseaux de neurones profonds (DNN) et récurrents (RNN) sont comparés, ainsi que l'utilisation d'un système de RAP pour détecter la frontière des mots. Les expériences présentées ici ont été conduites sur les corpus issus des campagnes d'évaluation ESTER1 et 2, ETAPE et REPERE1, constitués d'émissions de radio et de télévision française.

ABSTRACT

**Impact of speech activity detection systems for different automatic speech processing tasks**

In this article, we compare several Speech Activity Detection (SAD) systems and how they interact with two other speech processing tasks : Speaker Diarization and Automatic Speech Recognition (ASR). We study systems based on Gaussian Mixture Models (HMM/GMM), Deep and Recurrent Neural Networks (DNN and RNN) as well as an ASR system used to detect word borders. Experiments were made on French television and radio corpora : ESTER 1 and 2, ETAPE and REPERE1.

MOTS-CLÉS : détection de la parole, segmentation et regroupement de locuteurs, transcription automatique de la parole, apprentissage profond.

KEYWORDS: speech activity detection, speaker diarization, automatic speech recognition, deep learning.

## 1 Introduction

La détection de la parole est une étape importante dans plusieurs tâches de traitement de la parole. Elle permet d'extraire les segments de parole du signal, tout en ignorant le bruit, la musique, le silence, etc. Le but est d'une part de ne garder que des informations pertinentes pour la modélisation de la parole ou du locuteur et d'autre part de réduire la quantité de calculs nécessaire pour les traitements suivants.

De nombreuses approches permettent la détection de la parole. Il peut s'agir d'un seuillage sur l'énergie du signal ou d'autres paramètres acoustiques (Renevey & Drygajlo, 2001), d'une segmentation en utilisant des modèles de Markov cachés (*Hidden Markov Models*, HMM)(Kingsbury *et al.*, 2002; Gauvain *et al.*, 2002), de réseaux de neurones profonds (*Deep Neural Networks*, DNN)(Ryant *et al.*, 2013) ou récurrents (RNN)(Hughes & Mierle, 2013). Des mixtures de Gaussiennes (GMM) associées à des HMM sont souvent utilisées grâce au compromis qu'elles offrent entre légèreté et


10.21437/JEP.2018-63

| Statistiques | ESTER 1 | ESTER 2 | ETAPE | REPERE 1 |
|---|---|---|---|---|
| Nature | radio | radio | radio+TV | TV |
| Parole spontanée | peu | peu | beaucoup | variable |
| # de stations | 6 | 4 | 3 | 2 |
| # d'émissions | 9 | 8 | 11 | 7 |
| # d'enregistrements | 18 | 26 | 15 | 28 |
| Durée totale | 10h | 7h | 8h30 | 14h |
| Durée annotée | 9h | 6h | 6h | 3h |
| # de locuteurs uniques | 342 | 250 | 148 | 158 |
| Moyenne loc./émission | 20,17 | 11,53 | 10,33 | 7,5 |

TABLE 1 – Description du contenu des corpus évalués

performances. Cependant, ces dernières années, les performances des modèles neuronaux se sont grandement améliorées (Shahsavari *et al.*, 2017; Kaur & Sohal, 2017; Gelly & Gauvain, 2018).

En général, les systèmes de SAD servent comme première étape pour à d'autres tâches de traitement de la parole mais sont rarement optimisés pour celles-ci. Ce papier à pour but de comparer plusieurs système de détection de la parole afin de déterminer si l'un d'entre eux serais optimal pour la segmentation et le regroupement de locuteurs et la reconnaissance automatique de la parole.

Nous proposons de comparer plusieurs approches de détection de la parole en les utilisant pour des tâches de Segmentation et Regroupement de Locuteurs (SRL) et de Reconnaissance Automatique de la Parole (RAP). Nous utiliserons une approche GMM-HMM, trois approches basées sur les réseaux de neurones (récurrent et profond) et une approche utilisant un système de RAP.

Les différents corpus utilisés seront détaillés en Section 2 et les différents systèmes de détection de la parole en Section 3. La section 4 décrit les conditions d'évaluation (systèmes et métriques), tandis que les les résultats de ces comparaisons sont exposés en Section 5.

# 2 Corpus

Les corpus utilisés sont issus d'émissions de radio et de télévision française, il s'agit des corpus utilisés lors des campagnes d'évaluation ESTER1 (Gravier *et al.*, 2004), ESTER 2 (Galliano *et al.*, 2009), ETAPE (Gravier *et al.*, 2012) et REPERE 1 (Giraudel *et al.*, 2012). Leurs caractéristiques sont résumées dans le tableau 1. L'évaluation des différents systèmes se fait sur 87 enregistrements de journaux, de débats politiques ou d'émissions culturelles contenant des conditions acoustiques très difficiles.

Le corpus d'entraînement ainsi que le corpus de développement d'ESTER1 sont utilisés pour l'apprentissage des systèmes de détection de la parole. Le corpus d'apprentissage est partiellement annoté en 8 classes comme décrit dans (Meignier & Merlin, 2010). Les 8 classes consistent en 2 classes pour les silences en studio et au téléphone, 4 classes pour la parole en studio (propre , bruitée, avec de la musique ou d'un autre type), une classe pour la parole téléphonique et une classe pour les jingles et la musique pure. L'annotation a été réalisée de manière semi-automatique à partir d'un alignement forcé de la transcription de référence. Le tableau 2 indique les durées de chaque classe.

| Classe | Durée |
|---|---|
| Silence studio | 80 min |
| Silence téléphone | 35 min |
| Parole propre (F0, F1) | 54 min |
| Parole téléphonique (F2) | 62 min |
| Parole bruitée (F4) | 10 min |
| Parole et musique (F3) | 38 min |
| Parole autre (FX) | 142min |
| Jingles | 85min |

TABLE 2 – Description du corpus d'apprentissage Parole/Silence/Musique

# 3 Détection de la parole

## 3.1 Définition de la tâche

Le but d'un système de détection de la parole est de différencier les zones de parole des zones de non-parole au sein d'un signal audio, afin d'utiliser uniquement les segments de parole dans un système de reconnaissance du locuteur, de reconnaissance de la parole ou d'autres applications similaires.

La notion de non-parole inclut généralement tout ce qui est silence, bruit et musique, mais sa définition peut être élargie ou réduite selon la tâche. La notion de parole est aussi dépendante de la tâche : on peut vouloir garder la parole bruitée, superposée à de la musique, ou conserver uniquement la parole sans nuisances. Supprimer les segments de non-parole du signal permet d'éviter d'influencer les modèles d'apprentissage avec de l'information non pertinente et de réduire la quantité de calcul nécessaire à la tâche visée. Modéliser les caractéristiques de la non-parole est difficile, les sources de nuisances étant potentiellement nombreuses et très variables. La parole a moins de diversité et de variabilité, ce qui la rend comparativement plus simple à modéliser.

## 3.2 GMM-HMM

Le premier système de détection de la parole utilise des mixtures de gaussiennes (*Gaussian Mixture Models*, GMM) comme décrit dans (Meignier & Merlin, 2010). Il est réalisé à l'aide de la plateforme SIDEKIT (Larcher *et al.*, 2016) et de son extension S4D (*SIDEKIT for Diarization* [1]).

Le système est composé d'un modèle de Markov caché à 8 états, chacun d'eux associé à un GMM à 16 composantes diagonales. Ces 8 états/GMM représentent une classe acoustique du corpus d'apprentissage, représentant des états de la parole.

Les GMMs associés à l'HMM sont entraînés par l'algorithme EM-ML. En test, un décodage de Viterbi est utilisé pour retirer les zones de non-parole et garder les autres segments. Les pénalités de transmission entre les états ont été déterminées expérimentalement à partir d'un corpus de développement (campagne ESTER1).

Un post-traitement est nécessaire à la suite du décodage de Viterbi : le début et la fin de chaque segment de parole sont étendus de 0,5s afin de minimiser les imprécisions du décodage. Les segments de non-parole d'une durée inférieure à 0,25s sont réaffectés en parole.

1. https ://projets-lium.univ-lemans.fr/sidekit/

## 3.3 DNN-HMM

De façon similaire au GMM-HMM, le système se basant sur un réseau de neurones profond, est entraîné à classifier un vecteur acoustique en l'une des 8 classes du corpus. Les probabilités a posteriori générées par le DNN sont utilisées dans un décodage en Viterbi pour obtenir les segments de parole. L'étape de post-traitement est effectuée de la même manière que pour le système GMM-HMM. À la différence du GMM-HMM où 8 états sont décodés, ce système n'en utilise que 3 : les distributions de probabilités sont normalisées avant le décodage pour ne garder que la parole, le silence et la musique.

Le DNN acoustique est développé avec SIDEKIT et Theano (Bergstra *et al.*, 2010). Il est composé de 4 couches de 1000 neurones activées par une *sigmoïde*, et d'une couche à 8 sorties activée par un *Softmax*.

## 3.4 DNN « mimic »

Inspiré des travaux de (Rohdin *et al.*, 2017), ce système consiste à entraîner un DNN à reproduire les sorties du GMM utilisé en 3.2. L'hypothèse étant que le DNN a une meilleure capacité de généralisation que le GMM. Le DNN est construit à partir de Keras (Chollet, 2016) et est composé de 2 couches de 600 neurones activées par une fonction `tanh`, et d'une couche de 8 neurones à activation linéaire (les sorties du GMM étant des log-probabilités). L'erreur quadratique moyenne (*Mean Squared Error*, MSE) est utilisée en tant que fonction de coût pour l'apprentissage du réseau.

De la même façon que pour le GMM-HMM, un décodage de Viterbi et un post-traitement sont effectués afin de ne garder que les segments de parole.

## 3.5 Segmentation par transcription automatique

Un système de transcription automatique de la parole génère des mots ainsi que leurs frontières. L'idée est d'utiliser ces frontières comme segments de parole et de les évaluer tels quels.

Ce système est initialisé avec les segments du GMM-HMM mentionné précédemment. Le temps de calcul étant en fonction du nombre de locuteurs, un regroupement hiérarchique est effectué au préalable (décrit plus loin dans cet article). Les sorties du système de RAP sont traitées pour ne garder que l'information sur les frontières de mots, tout en supprimant les phonèmes trop longs considérer comme des erreurs de transcription.

## 3.6 LSTM bidirectionnel

Les LSTM (*Long Short-Term Memory* (Sepp Hochreiter & Jürgen Schmidhuber, 1997)) sont des réseaux de neurones récurrents possédant une mémoire interne à court et long terme. Le mode bidirectionnel permet la prise en compte d'un contexte acoustique plus large. Ce système permet de s'affranchir du décodage de Viterbi.

L'architecture utilisée s'inspire de (Yin *et al.*, 2017) mais adaptée à la détection de la parole et réalisée avec Keras. Il s'agit de deux couches de BLSTM à 64 et 40 neurones en sortie, deux couches de DNN à 40 et 10 neurones et d'une sortie à un neurone. Le réseau utilise un contexte long de 3,2s extrait toutes les 0.8s (avec un chevauchement de 75%). Le réseau produisant des segments du même format qu'en entrée, ceux-ci sont regroupés (moyenne) afin d'avoir un vecteur de probabilités par

fichier audio. Les segments de parole sont générés en utilisant deux seuils (activation/désactivation) sur les probabilités a posteriori du réseau.

L'apprentissage de ce système nécessite d'avoir une annotation continue du corpus utilisé pour l'apprentissage. Or le corpus d'apprentissage n'est que partiellement annoté. Pour ce système, les annotations utilisées sont les segments de parole pour la classe positive et les écarts entre ces segments pour la classe négative (non-parole).

# 4 Évaluation

## 4.1 Tâches évaluées

Les différents systèmes de détection de la parole ont été évalués dans deux tâches de traitement de la parole : la Segmentation et le Regroupement de Locuteurs (SRL), et la transcription automatique de la parole.

### 4.1.1 Segmentation et Regroupement de locuteurs

La tâche de segmentation et du regroupement de locuteurs consiste à répondre à la question « qui parle quand ? » dans un fichier audio. Typiquement, un système de SRL comporte quatre étapes : la paramétrisation, la détection de la parole, la segmentation en tour de parole, et le regroupement des tours par locuteurs uniques. L'étape de détection de la parole permet de supprimer les silences, bruits et musique qui dégraderaient la qualité des étapes suivantes. La segmentation en tour de parole a pour but de construire des segments de parole homogène et pure contenant un seul locuteur, pour enfin effectuer les regroupements (ou *clustering*) jusqu'à obtenir des classes représentant les locuteurs de l'enregistrement.

Le système utilisé ici est développé au LIUM, proche de celui décrit par (Dupuy *et al.*, 2014), mais réalisé à partir de l'extension S4D de SIDEKIT. Dans un premier temps, une segmentation acoustique est effectuée en utilisant le critère d'information bayésien (*Bayesian Information Criteria*, BIC), suivi par un regroupement hiérarchique utilisant la même métrique (*Hierarchical Agglomerative Clustering*, BIC-HAC) et un décodage de Viterbi, afin de réajuster les frontières des segments. À ce stade, généralement, chaque locuteur est représenté par plusieurs classes. Une dernière étape de regroupements est réalisée. Des *i*-vectors (Dehak *et al.*, 2011) sont extraits à partir de chaque classe, et une matrice de distance est calculée. Les regroupements sont affinés en traitant cette matrice comme un problème de Programmation Linéaire en Nombres Entiers (PLNE).

### 4.1.2 Tâche : Reconnaissance automatique de la parole

La tâche de Reconnaissance Automatique de la Parole (RAP) permet de générer le texte énoncé dans un fichier audio. Le système de transcription utilisé est similaire au système développé par le LIUM (Rousseau *et al.*, 2014) durant la campagne REPERE à partir de l'outil Kaldi (Povey *et al.*, 2011).

Le système de RAP prend en entrée chacune des segmentations générées par les systèmes de détection de la parole après l'étape de regroupement hiérarchique BIC décrite en 4.1.1.

## 4.2 Paramètres acoustiques pour les détecteurs de parole

Tous les systèmes évalués utilisent des MFCC (*Mel Frequency Cepstral Coefficients*). Les vecteurs sont extraits tous les 10ms sur une fenêtre de 25ms. Pour les systèmes GMM-HMM et MIMIC, 13 coefficients sans l'énergie sont extraits auxquels s'ajoute leurs dérivées (dimension 26). Pour le DNN-HMM, un contexte de $\pm 31$ trames (31 + 1 + 31 trames $\times$ 26 = 1638 coefficients) est ajouté.

Le BLSTM utilise une paramétrisation différente : 11 coefficients MFCC sont extraits de la même façon, auxquels sont ajoutées leurs dérivées premières et secondes avec leur énergie respective (11 + 12 + 12 = 35 coefficients).

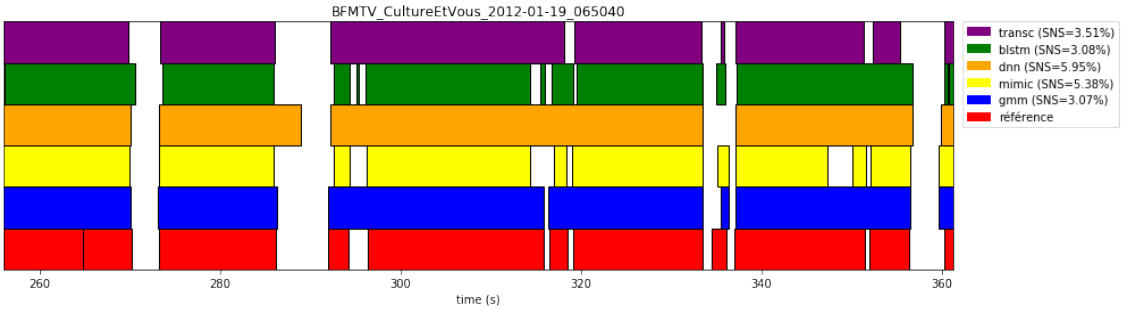
## 4.3 Métriques d'évaluation

Un système de détection de la parole est évalué à l'aide de deux scores : le taux de fausse alarme (FA), où le système détecte de la parole là où la référence n'en indique pas, et le taux de parole manquée (Miss), où le système échoue à détecter de la parole présente dans la référence. Ces deux taux peuvent être sommés pour mesurer le taux d'erreur parole vs non-parole (SNS).

Pour le regroupement de locuteurs, la métrique la plus utilisée est le DER (*Diarization Error Rate*), qui est composé de trois types d'erreurs : détection manquée, fausse détection et substitution de locuteurs par rapport à la référence. En complément au DER, les mesures de Pureté et de Couverture des segments sont utilisées comme décrites dans la documentation de l'outil utilisé *pyannote.metrics* (Bredin, 2017).

Enfin, pour la transcription, la métrique utilisée est le taux d'erreur de mots (*Word Error Rate*, WER), qui combine les erreurs de substitution, de suppression et d'insertion de mots.

FIGURE 1 – Exemple de segments générés




# 5 Résultats

La figure 1 montre les segments générés par les différents systèmes de SAD pour un fichier du corpus REPERE. On remarque que tous les systèmes réussissent à détecter les longs silences. Le DNN ne parvient pas à détecter les silences de courte durée et produit des segments de parole trop longs. Au contraire, le système utilisant la RAP et le système MIMIC génèrent des silences trop longs ou inexistants dans la référence, produisant un taux de détection manquée élevée (voir 3).

Le tableau 3 résume les résultats obtenus pour les 5 systèmes évalués. Au niveau de la tâche de SRL, le système GMM-HMM reste le meilleur système avec $0.24\%$ de DER d'avance en absolu par rapport au deuxième meilleur système.

Le DNN-HMM produit le taux de parole manquée le plus faible, mais aussi le plus haut taux de fausse alarme : les segments générés sont plus longs qu'ils ne devraient l'être. En conséquence ce système est le moins performant en termes de DER.

Le système « MIMIC » joue bien son rôle avec des taux d'erreur proches de ceux du GMM-HMM. Les résultats proposés sont ceux de la cinquième itération d'apprentissage du réseau.

Logiquement, le système utilisant la segmentation du système de RAP produit le plus faible taux de fausse alarme du lot grâce aux alignements qu'il effectue. Cependant le taux de parole manquée est aussi le plus élevé : certains éléments de la parole ne sont pas détectés comme des mots donc supprimés. Le DER qui en résulte est néanmoins le deuxième meilleur du groupe, mais ce système est pénalisé par un temps de calcul élevé par rapport aux autres systèmes.

Enfin, le LSTM bidirectionnel offre des taux d'erreur similaires au GMM-HMM, mais une augmentation $0.1\%$ absolue du taux d'erreur total se traduit par une augmentation nette de $0.55\%$ DER, classant ce système en avant-dernière position.

Quant aux différents taux d'erreur de mots, ceux-ci ne varient que très peu. Nous pouvons tout de même constater qu'il est préférable d'avoir un faible taux de parole manquée pour avoir un meilleur WER.

Un SAD « parfait » produirait un taux DER de $7.93\%$ et un WER de $14.00\%$. Il reste donc une marge d'amélioration pour ces systèmes, bien que celle-ci soit faible.

Le corpus d'évaluation, composé de 87 émissions très diverses, permet d'obtenir une performance moyenne des systèmes. Cependant le système GMM-HMM est devancé légèrement par la majorité des autres systèmes sur le corpus ESTER1 pour la tâche de SRL. Ce corpus est le corpus le plus facile en moyenne pour cette tâche. Il contient peu de parole spontanée et de parole superposée.

# 6 Conclusion

Cette étude conduite à partir d'un même corpus d'apprentissage nous a permis de comparer plusieurs systèmes de détection de la parole sur deux tâches de traitement de la parole : la segmentation et regroupement en locuteurs et la transcription. Le système classique GMM-HMM s'est avéré être le plus robuste grâce à son apprentissage de plusieurs modèles plutôt qu'un simple parole/non-parole binaire sur les deux tâches évaluées.

Sur les 87 enregistrements constituant un large corpus de test contenant des enregistrements nature divers, nous n'avons pas pu démontrer que les systèmes à base DNN surpassaient un système classique GMM-HMM. Il est à noter toutefois que le corpus d'apprentissage, construit au départ pour l'apprentissage du GMM-HMM développé durant la campagne ESTER1, n'a pas été remis en question.

Dans la mouvance actuelle qui consiste à développer des systèmes entièrement neuronaux (end-to-end), il faut noter que les performances des systèmes neuronaux, sans dépasser le système GMM-HMM, offrent des perspectives importantes. En effet, les systèmes présentés dans cette étude sont optimisés pour une tâche de détection de parole qui ne représente pas un cas d'usage mais seulement une première étape nécessaire aux systèmes finaux qui apportent une valeur ajoutée. L'intégration des systèmes de détection de parole neuronaux au sein d'une architecture complètement neuronale permettrait d'optimiser cette étape de sélection directement pour la tâche visée. Cette intégration constitue la prochaine étape de nos travaux.

| Systèmes | Corpus | FA | MISS | SNS | DER | Pureté | Couv. | WER |
|---|---|---|---|---|---|---|---|---|
| GMM-HMM | ESTER1 | 0.59 | 0.35 | 0.94 | 6.50 | 94.18 | 96.90 | 11.07 |
| | ESTER2 | 1.11 | 0.20 | 1.30 | 6.26 | 96.57 | 97.76 | 11.59 |
| | ETAPE | 3.19 | 0.24 | 3.43 | 15.69 | 84.30 | 85.29 | 25.96 |
| | REPERE | 0.53 | 0.95 | 1.48 | 9.30 | 88.74 | 91.25 | 14.89 |
| | Moyenne | 1.39 | 0.35 | **1.74** | **8.95** | 91.44 | 93.34 | **15.02** |
| DNN-HMM | ESTER1 | 0.67 | 0.30 | 0.97 | 5.85 | 94.34 | 96.99 | 11.21 |
| | ESTER2 | 1.93 | 0.47 | 2.40 | 7.91 | 96.83 | 97.76 | 11.99 |
| | ETAPE | 3.68 | 0.17 | 3.86 | 17.49 | 83.41 | 83.72 | 26.15 |
| | REPERE | 0.80 | 0.42 | 1.23 | 10.64 | 86.56 | 90.53 | 14.49 |
| | Moyenne | 1.80 | **0.33** | 2.13 | 9.71 | 90.70 | 92.85 | 15.07 |
| MIMIC | ESTER1 | 0.45 | 0.50 | 0.95 | 6.41 | 94.25 | 97.00 | 11.12 |
| | ESTER2 | 0.69 | 0.31 | 1.01 | 5.89 | 96.94 | 97.79 | 11.65 |
| | ETAPE | 2.86 | 0.46 | 3.33 | 17.10 | 83.46 | 83.69 | 26.16 |
| | REPERE | 0.41 | 2.04 | 2.46 | 10.50 | 87.76 | 90.76 | 16.11 |
| | Moyenne | 1.13 | 0.61 | **1.74** | 9.29 | 91.10 | 92.93 | 15.48 |
| TRANSCR. | ESTER1 | 0.33 | 0.89 | 1.22 | 5.66 | 96.03 | 96.90 | - |
| | ESTER2 | 0.47 | 0.53 | 1.01 | 6.19 | 97.05 | 97.46 | - |
| | ETAPE | 2.52 | 0.74 | 3.25 | 17.65 | 83.80 | 84.94 | - |
| | REPERE | 0.42 | 1.32 | 1.74 | 10.29 | 89.33 | 90.86 | - |
| | Moyenne | **0.94** | 0.80 | **1.74** | 9.19 | **92.07** | 93.06 | - |
| BLSTM | ESTER1 | 0.59 | 0.32 | 0.91 | 6.55 | 94.43 | 96.34 | 11.08 |
| | ESTER2 | 1.26 | 0.32 | 1.58 | 6.59 | 96.81 | 97.66 | 12.00 |
| | ETAPE | 3.15 | 0.31 | 3.46 | 17.20 | 82.68 | 85.62 | 26.21 |
| | REPERE | 0.49 | 1.29 | 1.78 | 10.09 | 87.62 | 91.90 | 15.80 |
| | Moyenne | 1.42 | 0.42 | 1.84 | 9.50 | 90.92 | **93.46** | 15.48 |

TABLE 3 – Résultats des systèmes de détection de parole appliqués en segmentation et regroupement et locuteur et en transcription automatique.
FA : détection de la parole erronée, MISS : détection manquée de parole, SNS : FA+MISS.
DER : erreur de segmentation et de regroupement en locuteur, Pureté et Couv. : pureté et couverture en locuteur des segments. WER : erreur de transcription automatique.

# Remerciements

Nous tenons à remercier Kong Aik Lee et Rafael E. Banchs de l'*Institute for Infocomm Research* ($I^2R$) ainsi qu'Antoine Laurent du LIUM pour leur aide précieuse.

# Références

BERGSTRA J., BREULEUX O., BASTIEN F., LAMBLIN P., PASCANU R., DESJARDINS G., TURIAN J., WARDE-FARLEY D. & BENGIO Y. (2010). Theano : A CPU and GPU math compiler in Python. In *Proc. 9th Python in Science Conf*, p. 1–7.

BREDIN H. (2017). pyannote. metrics : a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*.

CHOLLET F. (2016). keras : Deep Learning for humans. original-date : 2015-03-28T00 :35 :42Z.

DEHAK N., KENNY P. J., DEHAK R., DUMOUCHEL P. & OUELLET P. (2011). Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*.

DUPUY G., MEIGNIER S., DELÉGLISE P. & ESTEVE Y. (2014). Recent improvements on ilp-based clustering for broadcast news speaker diarization. In *Odyssey 2014*.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Tenth Annual Conference of the International Speech Communication Association*.

GAUVAIN J.-L., LAMEL L. & ADDA G. (2002). The LIMSI Broadcast News transcription system. *Speech Communication*, **37**(1), 89–108.

GELLY G. & GAUVAIN J. L. (2018). Optimization of RNN-Based Speech Activity Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **26**(3), 646–656.

GIRAUDEL A., CARRÉ M., MAPELLI V., KAHN J., GALIBERT O. & QUINTARD L. (2012). The REPERE Corpus : a multimodal corpus for person recognition. In *LREC*, p. 1102–1107.

GRAVIER G., ADDA G., PAULSON N., CARRÉ M., GIRAUDEL A. & GALIBERT O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *LREC-Eighth international conference on Language Resources and Evaluation*, p.ña.

GRAVIER G., BONASTRE J.-F., GEOFFROIS E., GALLIANO S., MCTAIT K. & CHOUKRI K. (2004). The ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. In *LREC*.

HUGHES T. & MIERLE K. (2013). Recurrent neural networks for voice activity detection. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* : IEEE.

KAUR S. & SOHAL J. S. (2017). Speech Activity Detection and its Evaluation in Speaker Diarization System. *INTERNATIONAL JOURNAL*, **16**(1).

KINGSBURY B., JAIN P. & ADAMI A. (2002). A hybrid HMM/TRAPS model for robust voice activity detection. In *Seventh International Conference on Spoken Language Processing*.

LARCHER A., LEE K. A. & MEIGNIER S. (2016). An extensible speaker identification sidekit in Python. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, p. 5095–5099 : IEEE.

MEIGNIER S. & MERLIN T. (2010). LIUM SpkDiarization : an open source toolkit for diarization. In *CMU SPUD Workshop*, volume 2010.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y. & SCHWARZ P. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding* : IEEE Signal Processing Society.

RENEVEY P. & DRYGAJLO A. (2001). Entropy based voice activity detection in very noisy conditions. p. 1887–1890.

ROHDIN J., SILNOVA A., DIEZ M., PLCHOT O., MATEJKA P. & BURGET L. (2017). End-to-end DNN Based Speaker Recognition Inspired by i-vector and PLDA. *arXiv :1710.02369 [cs, eess]*.

ROUSSEAU A., BOULIANNE G., DELÉGLISE P., ESTÈVE Y., GUPTA V. & MEIGNIER S. (2014). LIUM and CRIM ASR system combination for the REPERE evaluation campaign. In *International Conference on Text, Speech, and Dialogue*, p. 441–448 : Springer.

RYANT N., LIBERMAN M. & YUAN J. (2013). Speech activity detection on youtube using deep neural networks. In *INTERSPEECH*, p. 728–731.

SEPP HOCHREITER & JÜRGEN SCHMIDHUBER (1997). Long Short-Term Memory. *Neural Computation*, **9**(8), 1735–1780.

SHAHSAVARI S., SAMETI H. & HADIAN H. (2017). Speech activity detection using deep neural networks. In *2017 Iranian Conference on Electrical Engineering (ICEE)*, p. 1564–1568.

YIN R., BREDIN H. & BARRAS C. (2017). Speaker Change Detection in Broadcast TV Using Bidirectional Long Short-Term Memory Networks. p. 3827–3831 : ISCA.

XXXIIe Journées d'Études sur la Parole
4-8 juin 2018, Aix-en-Provence, France

# Impact des techniques d'adaptation au locuteur dans l'espace des paramètres pour des modèles acoustiques purement neuronaux

Natalia Tomashenko  Yannick Estève
LIUM, Le Mans Université, France
prenom.nom@univ-lemans.fr

## RÉSUMÉ

Cet article explore l'utilisation de techniques d'adaptation au locuteur pour des modèles acoustiques *bidirectionnels* de type *long short term memory* (BLSTM) entrainés avec la fonction objective dite de *classification temporelle connectionniste* (CTC). Les modèles acoustiques BLSTM-CTC prennent de plus en plus d'importance dans les systèmes de reconnaissance automatique de la parole, mais peu d'études ont été menées jusqu'ici pour y appliquer des techniques d'adaptation au locuteur. Dans cet article, nous explorons l'utilisation de trois techniques différentes : l'approche par *feature space maximum likelihood linear regression* (fMLLR), celle s'appuyant sur l'utilisation de i-vectors, et une approche exploitant la technique d'adaptation *maximum a posteriori* (MAP) appliquée sur des modèles gaussiens dont sont dérivés des paramètres fournis au modèles acoustiques neuronaux. Enfin, cette étude présente une comparaison du comportement des modèles BLSTM-CTC avec celui de modèles markoviens associés à un *time-delay neural network* (TDNN).

## ABSTRACT

**Exploration of feature-space speaker adaptation techniques for end-to-end acoustic models.**

This paper investigates speaker adaptation techniques for *bidirectional long short term memory* (BLSTM) recurrent neural network based acoustic models trained with the *connectionist temporal classification* (CTC) objective function. BLSTM-CTC AMs play an important role in end-to-end automatic speech recognition systems. However, there is a lack of research in speaker adaptation algorithms for these models. We explore three different feature-space adaptation approaches for CTC acoustic models : feature-space maximum linear regression, i-vector based adaptation, and maximum a posteriori adaptation using GMM-derived features. In addition, the adaptation behavior is compared for BLSTM-CTC models and time-delay neural network (TDNN) models trained with the cross-entropy criterion.

MOTS-CLÉS : Adaptation au locuteur, reconnaissance de la parole de bout en bout, paramètres acoustiques dérivés de GMM, réseaux de neurones profonds, modèles acoustiques.

KEYWORDS: Speaker adaptation, end-to-end speech recognition, GMM-derived features, deep neural network, acoustic model.



10.21437/JEP.2018-64

# 1 Introduction

Plusieurs approches neuronales de type bout en bout (*end-to-end*) ont récemment été proposées dans la littérature pour la reconnaissance automatique de la parole (Hannun *et al.*, 2014; Bahdanau *et al.*, 2016). Les modèles acoustiques (AMs) de type bout en bout tentent de produire des séquences de phonèmes ou de graphèmes à partir du signal de parole à l'aide d'architectures purement neuronales (Chorowski *et al.*, 2014; Graves & Jaitly, 2014; Miao *et al.*, 2015). Ils présentent une alternative à l'approche hybride devenue classique qui associe modèles de Markov cachés et réseaux de neurones profonds (HMM-DNNs).

L'adaptation au locuteur est un composant essentiel des modèles acoustiques hybrides HMM-DNNs à l'état de l'art, et plusieurs techniques d'adaptation ont été proposées pour les DNNs. Ces techniques comprennent la transformation linéaire, qui peut être appliquée à différents niveaux d'un système HMM-DNN (Gemello *et al.*, 2006), les techniques de régularisation, comme la régularisation L2-prior (Liao, 2013) ou la régularisation par divergence de Kullback-Leibler (Yu *et al.*, 2013), l'adaptation de l'espace du modèle (Swietojanski & Renals, 2014), l'apprentissage multitâche (Price *et al.*, 2014), l'adaptation factorisée (Li *et al.*, 2014), l'adaptation par codes de locuteurs (Xue *et al.*, 2014), l'utilisation de paramètres auxiliaires, comme les i-vecteurs (Saon *et al.*, 2013), les paramètres acoustiques dérivés de mélanges de modèles gaussiens (GMMD) (Tomashenko & Khokhlov, 2014), et bien d'autres. Pourtant, la majorité des travaux publiés au sujet des modèles acoustiques de bout en bout n'utilise pas de techniques d'adaptation au locuteur.

L'objectif de cet article est d'étudier l'impact de cette adaptation lorsqu'elle est utilisée pour des modèles acoustiques neuronaux de bout en bout. Dans notre étude, nous prenons l'exemple des modèles acoustiques *bidirectionnels* de type *long short term memory* (BLSTM) entrainés avec la fonction objective dite de *classification temporelle connectionniste* (CTC). Dans la suite de cet article nous nommerons ces modèles les modèles CTC. Pour évaluer cet impact, nous avons implémenté trois différentes techniques d'adaptation au locuteur pour ce type de modèles acoustiques, et avons mis en place une analyse expérimentale de ces méthodes. De plus, nous souhaitons comparer l'impact de ces techniques d'adaptation sur les modèles CTC à leur impact sur des modèles de Markov cachés associés à un *time-delay neural network* (TDNN) appris à l'aide du critère d'entropie croisée (CE).

La suite de l'article est organisée comme suit. Un rapide survol des modèles acoustiques neuronaux de bout en bout est présenté en section 2, et les résultats expérimentaux sont donnés en section 3. Enfin, une conclusion est fournie en section 4.

# 2 Reconnaissance de la parole neuronale de bout en bout

Une des premières avancées qui a permis de se rapprocher de systèmes de reconnaissance de la parole de type bout en bout a été proposée dans (Graves *et al.*, 2013) où, pour la tâche de reconnaissance de phonèmes, un réseau de neurones récurrent profond de type BLSTM est appris afin de transformer directement les séquences d'observations acoustiques en phonèmes. Cette proposition s'appuie sur la fonction de coût objective CTC (Graves *et al.*, 2006). C'est ce type de modèle CTC qui est utilisé dans notre étude. D'autres approches de type bout en bout ont également été présentées dans la littérature, comme les systèmes de type encodeur/décodeur avec mécanisme d'attention (Chorowski *et al.*, 2014; Bahdanau *et al.*, 2016), ou les réseaux de neurones convolutifs (Collobert *et al.*, 2016).

## 2.1 LSTMs bidirectionnels profonds

Les réseaux de neurones récurrents (RNNs) sont une extension des réseaux de neurones profonds (DNN) sur lesquels sont ajoutées des connections entre différents types d'unités neuronales, en particulier des connections vers des couches cachées d'états antérieurs. L'utilisation de la récurrence à travers la dimension temporelle permet aux RNNs de modéliser la dynamique du comportement d'un phénomène dans le temps. Afin de capturer l'information sur l'ensemble d'une séquence d'entrée, une architecture de réseau de neurones récurrent bidirectionnel (BRNN) a été proposée dans (Schuster & Paliwal, 1997). Dans les BRNNs, les données sont traitées dans deux directions (avant et arrière) à l'aide de deux couches cachées (une pour le traitement vers l'avant, l'autre vers l'arrière) qui alimentent la même couche de sortie. Les systèmes de reconnaissance de la parole à l'état de l'art utilisent des architectures neuronales profondes avec plusieurs couches cachées. Les sorties des couches cachées *forward* (vers l'avant) et *backward* (vers l'arrière) à l'instant $t$ sont concaténées : cette concaténation devient l'entrée des prochaines couches récurrentes. L'apprentissage des modèles RNN s'effectue généralement en appliquant l'algorithme d'apprentissage de rétropropagation à travers le temps (BPTT). Cependant, faire apprendre à des RNNs les dépendances temporelles longue distance peut devenir difficile en raison des problèmes de disparition et d'explosion du gradient (Bengio *et al.*, 1994). Pour éviter ce problème, les unités neuronales de type *long short-term memory* (LSTM) ont été introduites dans (Hochreiter & Schmidhuber, 1997). Dans le cadre de la reconnaissance automatique de la parole de bout en bout, les unités LSTM sont utilisées comme éléments de base des BRNNs (Miao *et al.*, 2015, 2016; Graves *et al.*, 2013).

## 2.2 Classification temporelle connectionniste

Avec l'approche CTC, l'alignement entre les éléments d'entrée et les étiquettes de sortie est inconnu. L'approche CTC peut être implémentée avec une couche de sortie de type *softmax* qui utilise une unité supplémentaire pour l'étiquette vide $\emptyset$. Le symbole $\emptyset$ correspond à l'émission d'aucune sortie et est utilisé pour estimer la probabilité de ne pas proposer d'étiquette de sortie à un instant donné. Le réseau de neurones est alors appris de manière à maximiser sur les données d'apprentissage la log-probabilité de toutes les séquences de sortie valides. L'ensemble des séquences valides d'étiquettes pour une séquence d'entrée est défini par l'ensemble de toutes les séquences possibles d'étiquettes telles que ces séquences soient construites dans le bon ordre, tout en acceptant les répétitions et l'étiquette $\emptyset$ entre deux étiquettes. Ces cibles pour l'apprentissage CTC peuvent être calculées en utilisant des transducteur a états finis (FSTs) et l'algorithme *forward-backward* peut être utilisé pour calculer la fonction de coût CTC. Aucune probabilité de transition d'états ou d'états initiaux n'est nécessaire pour l'approache CTC, au contraire de l'approche hybride DNN-HMM.

# 3 Résultats expérimentaux

Cette étude concerne principalement les techniques d'adaptation dans l'espace des paramètres pour les modèles neuronaux de bout en bout. Trois techniques d'adaptation des AMs ont été explorées dans nos expériences :

1. l'approche par fMLLR (Gales, 1998),
2. l'utilisation de i-vecteurs (Senior & Lopez-Moreno, 2014),
3. l'utilisation de l'adaptation MAP (Gauvain & Lee, 1994) appliquées à des paramètres GMMD (Tomashenko & Khokhlov, 2014, 2015; Tomashenko *et al.*, 2016b,c,a).

## 3.1 Données expérimentales

Les expériences ont été menées sur le corpus TED-LIUM (Rousseau *et al.*, 2014). Nous avons utilisé la dernière (seconde) version de ce corpus. Ce jeu de données publiquement disponible contient 1495 présentations des conférences TED, correspondant à 207 heures de parole pour 1242 locuteurs, enregistrées en 16kHz. Pour les expériences avec l'approche SAT (*speaker adaptative training*) et pour l'adaptation, nous avons retiré du corpus original les données correspondant à des locuteurs y apparaissant moins de 5 minutes. Le reste du corpus a été divisé en 4 parties : corpus d'apprentissage, corpus de développement, et deux corpus de test. Les caractéristiques générales des ensembles de données obtenus sont présentées dans la table 1. Pour l'évaluation, un modèle de langage 4-gram

| Caractéristiques | Train | Dev. | $\mathrm{Test}_1$ | $\mathrm{Test}_2$ |
|---|---|---|---|---|
| Durée totale, en heures | 171,66 | 3,49 | 3,49 | 4,90 |
| Durée moyenne par locuteur, en minutes | 10,0 | 15,0 | 15,0 | 21,0 |
| Nombre de locuteurs | 1 029 | 14 | 14 | 14 |
| Nombre de mots | - | 36 672 | 35 555 | 51 452 |

TABLE 1 – Statistiques du corpus de données.

avec un vocabulaire de 152k mots a été utilisé. Ce modèle de langage est proche du modèle "*small*" actuellement fourni dans la recette Kaldi *tedlium s5_r2*. Plus de détails sur ces données expérimentales sont fournis dans (Tomashenko *et al.*, 2016b).

## 3.2 Systèmes de référence (*baselines*)

Pour les expériences décrites dans cet article, nous avons utilisé la boîte à outils logicielle Kaldi (Povey *et al.*, 2011) et le système Eesen (Miao *et al.*, 2015), qui se distinguent principalement par leur modélisation acoustique (Eesen est dérivé de Kaldi).

Trois modèles acoustiques indépendants du locuteur ont été estimés avec le système Eesen, qui ne diffèrent que par les paramètres acoustiques utilisés pour représenter le signal de parole. Ces trois types de paramètres sont :

1. *fbanks* $\oplus\Delta \oplus \Delta\Delta$ *(dimension* $= 120$*)* : les paramètres de type bancs de filtre à 40 dimensions, concaténés avec leurs dérivées premières et secondes ;
2. les *paramètres MFCC à haute résolution (dimension* $= 40$*)* : ce sont des paramètres MFCC calculés sans réduction de dimension, en conservant les 40 cepstres ;
3. les paramètres de type *bottleneck (BN) (dimension* $= 40$*)*.

Le premier type de paramètres correspond à celui proposé dans la version originale de Eesen pour la recette liée au corpus TED-LIUM. Pour les modèles acoustiques des deux autres types de paramètres, nous avons appliqué deux stratégies d'augmentation des données sur les données d'apprentissage : perturbation de la vitesse (avec des facteurs 0,9 ; 1,0 et 1,1), et perturbation du volume comme proposé dans (Peddinti *et al.*, 2015).

Le premier modèle acoustique de référence a été appris de la manière décrite dans (Miao *et al.*, 2015), en utilisant le critère CTC et la même architecture BLSTM profonde. Le réseau de neurones BLSTM contient cinq couches bidirectionnelles d'unités BLSTM contenant 320 unités pour chaque

sous-couche *forward* (avant) ou *backward* (arrière). Les paramètres acoustiques, fournis en entrée du réseau de neurones, ont été normalisés à l'aide de la soustraction de la moyenne par locuteur, et à l'aide de la normalisation de la variance. La couche de sortie est une couche de type *softmax* à 41 dimensions qui correspondent à 39 phonèmes indépendants du contexte, un modèle de bruit et au symbole $\emptyset$.

Comme évoqué plus haut, le troisième modèle acoustique indépendant du locuteur a été appris sur des paramètres BN. Pour calculer ces paramètres, un modèle de type DNN (*feedforward Deep Neural Network*) a été construit avec l'architecture suivante : une couche d'entrée de 440 dimensions, quatre couches cachées à 1500 dimensions, sauf la troisième qui ne contient que 40 neurones et dont seront extraits les paramètres BN, et une couche de sortie de 4025 dimensions. Les paramètres acoustiques utilisés pour l'apprentissage de cet extracteur de *bottlenecks* sont le résultat de la concaténation (dimension par dimension : *splicing*) de 11 trames consécutives de paramètres MFCC de 40 dimensions chacune, pour un total de 440 paramètres.

## 3.3 Modèles acoustiques adaptés au locuteur

Trois techniques d'adaptation des modèles acoustiques ont été expérimentées de manière empirique dans cette section : fMLLR, adaptation par i-vecteur, et adaptation MAP appliquée sur des mélanges de modèles gaussiens dérivés. Les mêmes stratégies d'augmentation des données d'apprentissage telles qu'évoquées plus haut ont été appliquées pour tous les modèles acoustiques adaptés. Tous les modèles SAT ont été appris avec la même architecture neuronale (à l'exception de la couche d'entrée) et le même critère d'apprentissage, comme décrit dans la section 3.2 pour les modèles indépendants du locuteur (non adaptés). Les six modèles acoustiques SAT on été estimés avec les paramètres suivants :

4. *MFCC* $\oplus$ *i-vectors (dimension* $= 140$*)*;
5. *BN* $\oplus$ *i-vectors (dimension* $= 140$*)*;
6. *BN avec fMLLR (dimension* $= 40$*)*;
7. *MFCC* $\oplus$ *GMMD (dimension* $= 167$*)*;
8. *BN* $\oplus$ *GMMD (dimension* $= 167$*)*;
9. *BN avec fMLLR* $\oplus$ *GMMD (dimension* $= 167$*)*.

Pour les modèles acoustiques utilisant les paramètres #4 et #5, les i-vecteurs de 100 dimensions ont été calculés *online* comme présenté dans (Peddinti *et al.*, 2015), et les statistiques utilisées pour le calcul des i-vecteurs ont été mises à jour toutes les deux phrases durant l'apprentissage. Pour les modèles acoustiques de #7 à #9, nous avons appliqué les paramètres BN pour l'apprentissage du modèle auxiliaire de type GMM (modèle de mélange gaussien) utilisé pour l'extraction des paramètres GMMD. Les paramètres GMMD adaptés au locuteur ont été obtenus de la manière décrite dans (Tomashenko *et al.*, 2016b).

## 3.4 Résultats expérimentaux pour les modèles acoustiques de type CTC

Pour l'ensemble des résultats proposés dans cette section, les techniques d'adaptation ont été appliquées de manière non supervisée lors du décodage des données de test en utilisant les transcriptions automatiques générées par le meilleur modèle acoustique générique (non adapté). Les taux d'erreurs

sur les mots obtenus par les différents systèmes sont présentés dans le tableau 2. Les trois premières lignes du tableau (#1–#3) correspondent aux modèles acoustiques génériques de référence (*baselines*), qui ont été construits selon les approches décrites en section 3.2. La première ligne représente le système Eesen tel qu'il est distribué (Miao *et al.*, 2015). Les six lignes suivantes (#4–#9) présentent les résultats des modèles adaptés. La numérotation utilisée dans la table 2 coïncide avec la numérotation des sections 3.2 et 3.3. Les deux dernières lignes du tableau (#10 et #11) ont été obtenues en utilisant les modèles acoustiques des lignes #8 et #9, mais l'adaptation des paramètres GMMD (notés alors GMMD* dans les tableaux 2, 3) de #10 et #11 a été effectuée à partir des transcriptions obtenu à l'aide du modèle adapté #6, alors que pour toutes les autres expériences l'adaptation a été réalisée à partir des transcriptions obtenues grâce au modèle générique #2. Parmi les systèmes CTC #1–#9, c'est le système #9 qui obtient les meilleurs résultats. Ce système correspond à l'utilisation de paramètres GMMD adaptés par MAP, et concaténés avec des paramètres BN adaptés par fMLLR. Une légère amélioration (#11) peut être obtenue en réadaptant les paramètres du modèle à partir des transcriptions automatiques que ce modèle obtient dans une première passe. De toutes les techniques d'adaptation appliquées séparément (#4–#8), c'est l'adaptation des paramètres GMMD par MAP qui obtient les meilleures performances, que ce soit avec les paramètres BN ou les paramètres MFCC.

| # | Paramètres | **CTC** : WER,% | | | **TDNN** : WER,% | | |
|---|---|---|---|---|---|---|---|
| | | Dev. | $\text{Test}_1$ | $\text{Test}_2$ | Dev. | $\text{Test}_1$ | $\text{Test}_2$ |
| 1 | fbanks $\oplus\Delta \oplus \Delta\Delta$ | 14,57 | 11,71 | 15,29 | - | - | - |
| 2 | MFCC | 13,21 | 11,16 | 14,15 | 13,69 | 11,34 | 14,38 |
| 3 | BN | 13,63 | 11,84 | 15,06 | 12,32 | 10,48 | 14,00 |
| 4 | MFCC $\oplus$ i-vecteurs | 12,92 | 10,45 | 14,09 | 11,63 | 9,62 | 13,28 |
| 5 | BN $\oplus$ i-vecteurs | 13,47 | 11,37 | 14,31 | 11,62 | 9,75 | 13,30 |
| 6 | BN-fMLLR | 12,45 | 10,96 | 13,79 | **10,70** | **9,28** | **12,84** |
| 7 | MFCC $\oplus$ GMMD | 11,95 | 10,20 | 14,04 | 11,30 | 9,75 | 13,74 |
| 8 | BN $\oplus$ GMMD | 11,66 | 10,14 | 13,88 | 11,07 | 9,75 | 13,55 |
| 9 | BN-fMLLR $\oplus$ GMMD | **11,63** | **9,91** | **13,58** | 10,92 | 9,54 | 13,27 |
| 10 | BN $\oplus$ GMMD$^*$ | 11,67 | 10,11 | 13,70 | 10,29 | 9,20 | 13,04 |
| 11 | BN-fMLLR $\oplus$ GMMD$^*$ | **11,41** | 9,93 | **13,47** | **10,15** | **9,06** | **12,84** |

TABLE 2 – Taux d'erreurs sur les mots (WER) en fonction des paramètres acoustiques, de la technique d'adaptation (aucune de #1 à #3) et du type de modèles acoustiques (CTC ou TDNN). GMMD* correspond aux paramètres GMMD adaptés à partir des transcriptions produites par un modèle SAT (par défaut ces transcriptions proviennent de l'utilisation d'un modèle générique).

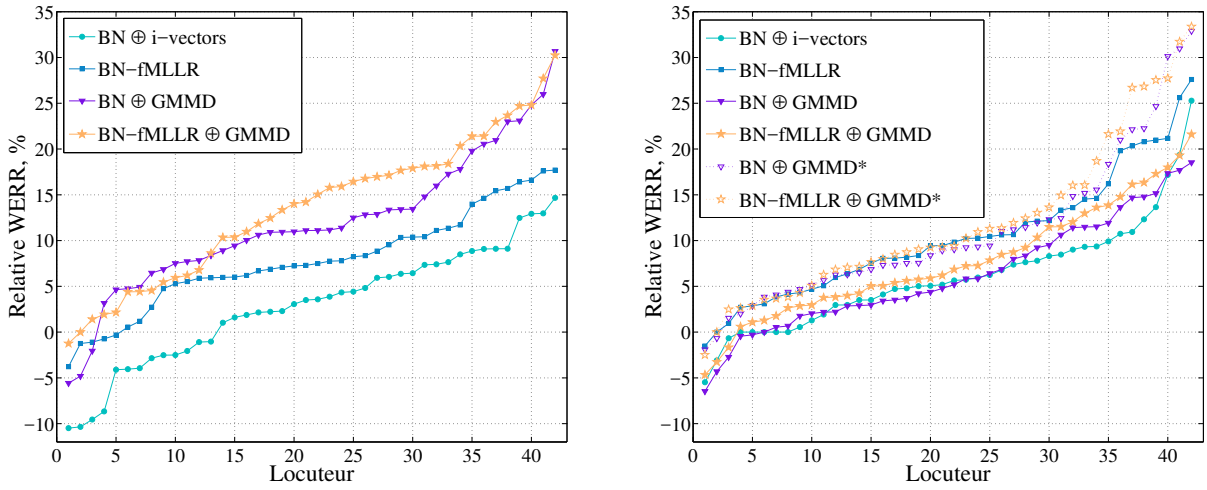
## 3.5 Comparaison de l'impact des techniques adaptation en fonction de la nature des modèles acoustiques : CTC ou TDNN

Dans cette section, nous souhaitons comparer le comportement des techniques d'adaptation lorsqu'elles sont appliquées à des modèles CTC à leur comportement lorsqu'elles sont appliquées à des modèles hybrides combinant modèles de Markov cachés et réseau de neurones. Pour cela, nous avons choisi de refaire les mêmes expériences que pour les modèles CTC en utilisant un modèle de type TDNN, car ces modèles sont actuellement souvent utilisés dans les systèmes à l'état de l'art (Peddinti *et al.*, 2015). Ces modèles acoustiques ont été estimés avec le critère d'entropie croisée (CE). Pour réaliser cette comparaison, nous avons construit le même ensemble de modèles acoustiques génériques

| # | Paramètres | **CTC** : rel. WERR,% | | | **TDNN** : rel. WERR,% | | |
|---|---|---|---|---|---|---|---|
| | | Dev. | $\text{Test}_1$ | $\text{Test}_2$ | Dev. | $\text{Test}_1$ | $\text{Test}_2$ |
| 4 | MFCC $\oplus$ i-vecteurs | 2,2 | 6,4 | 0,4 | 5,6 | 8,2 | 5,1 |
| 5 | BN $\oplus$ i-vecteurs | -2,0 | -1,9 | -1,1 | 5,7 | 7,0 | 5,0 |
| 6 | BN-fMLLR | 5,8 | 1,8 | 2,5 | 13,2 | 11,5 | 8,3 |
| 7 | MFCC $\oplus$ GMMD | 9,5 | 8,6 | 0,8 | 8,3 | 7,0 | 1,9 |
| 8 | BN $\oplus$ GMMD | 11,7 | 9,1 | 1,9 | 10,2 | 7,0 | 3,2 |
| 9 | BN-fMLLR $\oplus$ GMMD | 12,0 | **11,2** | 4,0 | 11,4 | 9,0 | 5,2 |
| 10 | BN $\oplus$ GMMD* | 11,7 | 9,4 | 3,2 | 16,5 | 12,2 | 6,9 |
| 11 | BN-fMLLR $\oplus$ GMMD* | **13,6** | 11,0 | **4,8** | **17,6** | **13,6** | **8,3** |

TABLE 3 – Réduction relative du taux d'erreurs sur les mots (WERR) pour les modèles acoustiques adaptés de type CTC et TDNN, par rapport au meilleur module générique (non adapté) de chaque type de modèle (#2 pour CTC et #3 pour TDNN). Ces valeurs sont calculées à partir des résultats du tableau 2.

et adaptés que nous avions construit pour les modèles CTC (voir sections 3.2 et 3.3), à l'exception du modèle #1. Tous les modèles TDNN ont été appris de la même manière et utilisent la même architecture neuronale. Ils ne diffèrent que par la nature des paramètres acoustiques qu'ils doivent traiter. L'architecture des modèles TDNN est semblable à celle décrite dans (Peddinti *et al.*, 2015). Le contexte temporel est de la forme $[t-16, t+12]$, alors que les index de concaténation (splicing) que nous avons utilisés sont les suivants : $[-2, 2]$, $\{-1, 2\}$, $\{-3, 3\}$, $\{-7, 2\}$, $\{0\}$, $\{0\}$. Ces modèles utilisent des couches cachés de 850 dimensions avec fonctions d'activation ReLU (Rectified Linear Units), et une couche de sortie d'environ 4000 dimensions.




(a) CTC (b) TDNN

FIGURE 1 – Réduction relative du taux d'erreurs sur les mots (WERR) pour les locuteurs des corpus de test et de développement en fonction de la technique d'adaptation utilisée, et comparativement au modèle acoustique utilisant des paramètres BN (#3). Pour chaque modèle acoustique, les résultats sont classés par ordre croissant du WERR.

Comme les résultats des modèles CTC, les résultats expérimentaux liés aux modèles TDNN sont

indiqués dans le tableau 2 et la figure 1. Pour les modèles TDNN, l'adaptation est effectuée à partir des transcriptions automatiques générées avec le modèle utilisant des paramètres BN. Dans la figure 1b, pour les modèles TDNN nous avons ajouté l'utilisation de modèles SAT (GMMD*) pour générer les transcriptions pour l'adaptation, car elle permet une amélioration plus conséquente des performances que ce que nous avions observé avec les modèles CTC. Le tableau 3 montre les réductions relatives du taux d'erreurs en fonction des techniques d'adaptation pour les modèles CTC et TDNN, par rapport au meilleur modèle acoustique générique correspondant (#2 pour CTC et #3 pour TDNN). Comme nous pouvons le constater, le choix des paramètres optimaux dépend de la nature du modèle acoustiques. Pour les modèles TDNN, nous observons dans nos expériences que les paramètres BN donnent de meilleurs résultats que les MFCC, alors que pour les modèles CTC la situation est inversée. De plus, nous remarquons que le classement des systèmes en fonction de leur taux d'erreurs diffère selon la nature (CTC ou TDNN) des modèles acoustiques.

# 4 Conclusions

Cet article porte sur l'étude du bénéfice potentiel apporté par les techniques d'adaptation au locuteur aux systèmes de reconnaissance de la parole neuronaux de bout en bout. Il montre que l'adaptation au locuteur reste un mécanisme essentiel pour l'amélioration des performances dans ce nouveau paradigme pour la reconnaissance de la parole. Les résultats expérimentaux sur le corpus TED-LIUM montrent que, dans un mode non supervisé, les techniques d'adaptation et d'augmentation des données d'apprentissage peuvent apporter une réduction relative du taux d'erreurs sur les mots comprise entre 10 et 20%, par exemple en se comparant à un modèle CTC générique utilisant des paramètres de type banc de filtres. En moyenne pour les modèles CTC, les meilleurs résultats sont obtenus en utilisant des paramètres dérivés de GMMs adaptés par MAP, en les combinant à des *bottlenecks* adaptés par fMLLR. Nous avons montré que les performances obtenues par les différentes techniques d'adaptation dépendent de la nature de l'architecture neuronale d'un modèle acoustique. Enfin, cet article présente des résultats expérimentaux qui permettent de comparer dans des conditions réalistes les performances des approches BLSTM-CTC à celles des approches à l'état de l'art comme les HMM-TDNN. Le taux d'erreurs sur les mots du meilleur modèle générique de type TDNN est relativement plus bas de 1 à 7% par rapport à celui du meilleur modèle CTC. Avec les modèles adaptés, cet écart augmente et le meilleur modèle TDNN commet entre 5 à 13% d'erreurs de moins que le meilleur modèle CTC.

# Références


BAHDANAU D., CHOROWSKI J., SERDYUK D., BRAKEL P. & BENGIO Y. (2016). End-to-end attention-based large vocabulary speech recognition. In *ICASSP*, p. 4945–4949 : IEEE.

BENGIO Y., SIMARD P. & FRASCONI P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, **5**(2), 157–166.

CHOROWSKI J., BAHDANAU D., CHO K. & BENGIO Y. (2014). End-to-end continuous speech recognition using attention-based recurrent NN : First results. *arXiv preprint arXiv :1412.1602*.

COLLOBERT R., PUHRSCH C. & SYNNAEVE G. (2016). Wav2letter : an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv :1609.03193*.

GALES M. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer speech and language*, **12**(2), 75–98.

GAUVAIN J.-L. & LEE C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains. *IEEE Trans. Speech and Audio Proc.*, **2**, 291–298.

GEMELLO R., MANA F., SCANZIO S., LAFACE P. & DE MORI R. (2006). Adaptation of hybrid ANN/HMM models using linear hidden transformations and conservative training. In *ICASSP*, p. 1189–1192.

GRAVES A. *et al.* (2006). Connectionist temporal classification : labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, p. 369–376 : ACM.

GRAVES A. & JAITLY N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In *ICML*, volume 14, p. 1764–1772.

GRAVES A., MOHAMED A.-R. & HINTON G. (2013). Speech recognition with deep recurrent neural networks. In *ICASSP*, p. 6645–6649 : IEEE.

HANNUN A., CASE C., CASPER J., CATANZARO B., DIAMOS G., ELSEN E., PRENGER R. *et al.* (2014). Deep speech : Scaling up end-to-end speech recognition. *arXiv preprint arXiv :1412.5567*.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long short-term memory. *Neural computation*, **9**(8), 1735–1780.

LI J., HUANG J.-T. & GONG Y. (2014). Factorized adaptation for deep neural network. In *ICASSP*, p. 5537–5541 : IEEE.

LIAO H. (2013). Speaker adaptation of context dependent deep neural networks. In *ICASSP*, p. 7947–7951.

MIAO Y. *et al.* (2016). An empirical exploration of CTC acoustic models. In *ICASSP*, p. 2623–2627 : IEEE.

MIAO Y., GOWAYYED M. & METZE F. (2015). Eesen : End-to-end speech recognition using deep rnn models and wfst-based decoding. In *ASRU*, p. 167–174 : IEEE.

PEDDINTI V., POVEY D. & KHUDANPUR S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *INTERSPEECH*, p. 3214–3218.

POVEY D. *et al.* (2011). The Kaldi speech recognition toolkit. In *ASRU*.

PRICE R., ISO K. & SHINODA K. (2014). Speaker adaptation of deep neural networks using a hierarchy of output layers. In *SLT*, p. 153–158 : IEEE.

ROUSSEAU A., DELÉGLISE P. & ESTÈVE Y. (2014). Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks. In *LREC*, p. 3935–3939.

SAON G., SOLTAU H., NAHAMOO D. & PICHENY M. (2013). Speaker adaptation of neural network acoustic models using i-vectors. In *ASRU*, p. 55–59.

SCHUSTER M. & PALIWAL K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, **45**(11), 2673–2681.

SENIOR A. & LOPEZ-MORENO I. (2014). Improving DNN speaker independence with i-vector inputs. In *ICASSP*, p. 225–229.

SWIETOJANSKI P. & RENALS S. (2014). Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models. In *SLT*, p. 171–176 : IEEE.

TOMASHENKO N. *et al.* (2016a). Exploration de paramètres acoustiques dérivés de GMMs pour l'adaptation non supervisée de modèles acoustiques à base de réseaux de neurones profonds. In *JEP*, p. 337–345.

TOMASHENKO N. & KHOKHLOV Y. (2014). Speaker adaptation of context dependent deep neural networks based on MAP-adaptation and GMM-derived feature processing. In *INTERSPEECH*, p. 2997–3001.

TOMASHENKO N. & KHOKHLOV Y. (2015). GMM-derived features for effective unsupervised adaptation of deep neural network acoustic models. In *INTERSPEECH*, p. 2882–2886.

TOMASHENKO N., KHOKHLOV Y. & ESTEVE Y. (2016b). On the use of Gaussian mixture model framework to improve speaker adaptation of deep neural network acoustic models. In *INTERSPEECH*, p. 3788–3792.

TOMASHENKO N., KHOKHLOV Y., LARCHER A. & ESTEVE Y. (2016c). Exploring GMM-derived features for unsupervised adaptation of deep neural network acoustic models. In *SPECOM*, p. 304–311.

XUE S. *et al.* (2014). Fast adaptation of deep neural network based on discriminant codes for speech recognition. *Audio, Speech, and Language Processing, IEEE/ACM Trans. on*, **22**(12), 1713–1725.

YU D., YAO K., SU H., LI G. & SEIDE F. (2013). KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In *ICASSP*, p. 7893–7897.

# Influence de la posture corporelle sur les paramètres acoustiques de la parole

Anaïs Delhoume, Emmanuel Ferragne

Laboratoire CLILLAC-ARP EA 3967, Université Paris Diderot, 75013 Paris

anais.delhoume@gmail.com, emmanuel.ferragne@univ-paris-diderot.fr

RESUME

Peu de recherches se sont intéressées à l'influence que peut exercer la posture corporelle sur la production de la parole et notamment sur ses paramètres acoustiques. Cette étude examine l'influence de la position allongée et de la position assise sur les formants (F1 et F2) des 10 voyelles orales du français, sur F0, sur le débit de parole et sur les caractéristiques du signal électroglottographique. Les faibles différences observées entre les deux conditions mettent en avant la capacité des locuteurs à compenser les effets de la gravité pour préserver les caractéristiques acoustiques de la parole.

ABSTRACT

**The influence of body posture on the acoustic parameters of speech**

Few studies have investigated the influence of body posture on speech production and especially on its acoustic parameters. This study examines the influence of supine and upright position on the formants (F1 and F2), of the 10 French oral vowels, on F0, on speech rate and on the characteristics of the electroglottographic signal. The small differences observed between the two conditions highlight the speakers' ability to compensate for the effects of gravity in an attempt to preserve the acoustic characteristics of speech.

MOTS-CLES : phonétique acoustique, voyelles orales, formants, position corporelle
KEYWORDS: acoustic phonetics, oral vowels, formants, body position

# 1 Introduction

Plusieurs études ont évalué l'impact de la posture sur les mouvements articulatoires de la langue (Kitamura et al., 2005 ; Stone et al., 2007 ; Tiede et al., 2000 ; Wrench et al., 2011 ; Traser et al., 2013), de la mâchoire (Tiede et al., 2000 ; Shiller et al., 1999 ; Traser et al., 2013) et des lèvres (Kitamura et al., 2005 ; Wrench et al., 2011 ; Traser et al., 2013). La gravité affecte la position de la langue et ses mouvements (Kitamura et al., 2005 ; Stone et al., 2007). En effet, la langue tend à se rétracter sous la force gravitationnelle en position allongée pour les voyelles postérieures. Cependant, le locuteur est capable de compenser ces effets (Kitamura et al., 2005 ; Stone et al., 2007). La position allongée semble limiter le degré d'ouverture de la mâchoire et les lèvres tendent à


10.21437/JEP.2018-65

être plus minces (Kitamura et al., 2005 ; Engwall, 2006). En position allongée, le larynx remonte (Kitamura et al., 2005 ; Traser et al., 2013), ce qui entraîne une modification des cavités orales et donc des fréquences de résonance. Cependant, l'impact de la gravité sur les caractéristiques articulatoires et la capacité des locuteurs à contrer ses effets varient d'un sujet à un autre et selon les voyelles à produire.

Seules quelques études se sont également intéressées aux conséquences d'un changement de posture corporelle sur les paramètres acoustiques de la parole. Elles ont toutes porté sur la langue anglaise, et les résultats issus de ces différentes recherches sont contradictoires. Tiede et al. (2000) n'ont pas constaté de différences dans les trois premiers formants ; ils suggèrent que la configuration articulatoire globale annule toutes les différences anatomiques spécifiques à chaque posture. Stone et al. (2007) n'ont pas non plus constaté de différences dans les fréquences formantiques selon la posture. Cependant, Shiller et al. (1999) ont mesuré des différences dans les formants F1 et F2 pour les voyelles /e/ et /æ/ : pour ces deux voyelles, les fréquences de F1 sont plus basses en position allongée et celles de F2 sont plus hautes. Hoedl (2015) suggère que la posture du locuteur influence l'articulation des voyelles et leurs paramètres acoustiques. Cependant, l'auteur indique que toutes les voyelles et tous les formants ne sont pas influencés de la même manière. Alors que F1 ne semble pas être affecté par les changements de posture, F2 et F3 subissent des modifications pour certaines voyelles lorsqu'elles sont produites en position allongée : pour F2, l'effet de la posture est significatif sur les voyelles /iː/, /e/ et /æ/ et non sur /ɔː/ et /uː/. Pour F3, la différence est significative pour les voyelles /iː/, /æ/ et /uː/ et elle ne l'est pas pour /e/ ni pour /ɔː/. D'après Hoedl, la position allongée entraîne une diminution des fréquences de F2 et F3.

F1 et F2 sont importants pour l'intelligibilité de la parole et devraient donc entraîner la mise en place de stratégies compensatoires selon Flory et Nolan (2015). Néanmoins, il convient de nuancer ces propos : en effet, la réalisation phonétique des voyelles est intrinsèquement variable. Nous pouvons par conséquent prédire qu'en deçà d'une déviation importante des formes canoniques, une compensation s'avère superflue. Flory et Nolan (2015) observent que F3 diffère régulièrement en position ventrale, démontrant ainsi que les sujets ne compensent pas complètement les changements que subit le conduit vocal à la suite du changement de position. Les valeurs de F3 augmentent en position ventrale pour la voyelle /æ/ et restent similaires en position dorsale et en station assise. Selon Flory et Nolan, F0 est probablement le paramètre acoustique qui varie le plus en raison de ses caractéristiques linguistiques et paralinguistiques : la fréquence fondamentale est significativement plus haute en position ventrale – en raison de l'augmentation de la tension dans les structures périlaryngées – et reste inchangée en position dorsale et en station assise du fait d'une tension moindre dans ces mêmes structures. Pour tous les formants, la position ventrale a abouti à un plus grand nombre de différences, en opposition à la station assise et à la position dorsale.

Il est donc intéressant d'étudier l'impact de la posture corporelle sur la production de la parole dans la mesure où celle-ci peut être produite en position allongée dans certaines situations spécifiques, notamment lors du sommeil : cette étude apparaît donc comme une base nécessaire et intéressante pour enrichir les recherches menées sur la somniloquie (Arnulf et al., 2017). En effet, certains sujets somniloques – notamment les personnes somnambules et celles atteintes de Trouble du

Comportement en Sommeil Paradoxal – sont fréquemment amenés à produire des vocalisations en positions assise ou allongée.

Dans cet article, nous employons les termes « position assise » et « position allongée » pour décrire la posture corporelle des locuteurs. L'orientation de la tête en position assise est non inclinée. Le terme « position allongée » renvoie à la posture qu'adopte le locuteur lorsqu'il est en décubitus dorsal avec la tête alignée dans l'axe du corps.

D'après les résultats issus des précédentes recherches, nous devrions observer une différence de la valeur des fréquences formantiques des voyelles orales du français consécutive au changement de position corporelle. En l'absence de compensation, F2 devrait diminuer en position allongée sous l'effet d'une articulation plus postérieure. Cette variation dans les valeurs formantiques devrait dépendre de la voyelle à produire et du formant analysé.

# 2 Expérience

## 2.1 Méthode

### *2.1.1 Sujets*

16 locuteurs natifs du français ont participé à l'étude : 12 femmes et 4 hommes âgés de 21 à 52 ans. Ils ont tous été enregistrés dans deux positions corporelles : assise et allongée. Un groupe de 9 personnes a commencé en station assise puis est passé en position allongée, et inversement pour l'autre groupe constitué de 7 personnes. Les enregistrements dans les deux conditions ont été réalisés consécutivement.

### *2.1.2. Stimuli*

Les sujets étaient invités à lire trois fois une liste de dix monosyllabes. Chaque monosyllabe contenait une voyelle orale précédée de la consonne /b/. Les items ainsi formés étaient les suivants : *ba, baie, bé, beau, beu, beurre, bi, botte, bou, bu*. Les listes contenaient les mêmes items mais ordonnés de trois façons différentes, dans un ordre aléatoire. Ces listes étaient accompagnées d'un extrait composé de 39 mots issu du livre *Le Petit Prince* d'Antoine De Saint-Exupéry dans lequel nous retrouvons les dix voyelles orales.

### *2.1.3. Procédure*

Le signal audio a été capturé au moyen d'un microphone à condensateur ECM8000 de Behringer placé à environ 15 centimètres des lèvres des participants et branché à un électroglottographe EG2 de Glottal Enterprises. Le signal EGG a été acquis simultanément à l'audio, et les deux types de signaux ont été numérisés en 16 bits avec un échantillonnage de 22050 Hz, conformément aux recommandations pour le matériel utilisé, chacun sur une piste d'un fichier PCM stéréo.

Les sujets ont été enregistrés en train de lire les trois listes de monosyllabes – à une vitesse de lecture d'environ un stimulus par seconde – puis le texte, dans les deux conditions. En position assise, les locuteurs étaient assis au bord d'un lit et en position allongée, ils avaient la tête légèrement surélevée sur un oreiller. La durée de la passation était d'approximativement quinze minutes.

*2.1.4. Analyses acoustique et statistique*

Nous avons effectué une analyse acoustique des signaux recueillis. Les voyelles ont été segmentées manuellement avec Praat en nous appuyant sur la présence des bandes formantiques dans le spectrogramme. L'estimation des formants a ensuite été conduite de façon semi-automatique en ajustant manuellement les paramètres de l'algorithme jusqu'à obtenir une superposition optimale des tracés estimés et des pics d'énergie sur le spectrogramme[1].

Les valeurs de F1 et F2 dans les analyses qui suivent ont été extraites au milieu temporel de la voyelle et exprimées sur l'échelle des Bark. En plus des mesures par voyelle, nous avons calculé la dispersion globale de l'espace vocalique de chaque locuteur dans chaque condition en estimant l'aire de l'enveloppe convexe de toutes les voyelles dans F1-F2. Nous avons, en outre, souhaité tester une potentielle différence de position de l'espace acoustique en calculant le F1 moyen et le F2 moyen de tout l'espace. Nous avons renoncé à inclure F3 dans nos analyses car les estimations se sont révélées peu fiables.

L'extrait de texte a été utilisé pour calculer les valeurs moyennes et les écarts-types de F0 en demi-tons ainsi que pour analyser le débit de parole, en syllabes par seconde. La durée des voyelles a également été calculée. Nous avons enfin mené une étude préliminaire sur l'analyse du signal EGG en mesurant le quotient ouvert moyen de la voyelle /a/ de chaque locuteur par condition. A ce stade exploratoire pour l'EGG, nous nous sommes contentés d'une estimation du quotient ouvert en utilisant un seuil correspondant à 35% de l'amplitude crête-à-crête de chaque période.

Nous avons appliqué des modèles linéaires mixtes, avec les facteurs fixes Position (assis vs allongé) et Voyelle (quand celui-ci s'applique), le facteur aléatoire Sujet, et, successivement, les variables dépendantes suivantes : les fréquences de F1 et F2 (Bark) calculées au milieu temporel, l'aire de l'enveloppe convexe englobant toutes les voyelles, les F1-F2 moyens de tout l'espace, la durée des voyelles, la moyenne et l'écart-type de F0 mesurés en demi-tons, le débit de parole en syllabes par seconde (tous deux à partir de l'extrait de texte) et le quotient ouvert (EGG) moyen sur la voyelle /a/.

# 3 Résultats

Un premier modèle avec les facteurs fixes Voyelle et Position, le facteur aléatoire Sujet et la variable dépendante (VD) F1 en Bark ne fait apparaître aucun effet significatif ni aucune interaction entre les deux facteurs fixes. Le même modèle, cette fois-ci avec F2, conduit à l'obtention de

[1]Scripts disponibles à cette adresse : https://tinyurl.com/hwv6a96

résultats similaires. Afin de tester un effet potentiel de la position du sujet sur les F1-F2 moyens de tout l'espace formantique, un modèle avec le facteur Position (fixe), le facteur Sujet (aléatoire) a été appliqué. Là encore, aucune différence significative n'émerge. Un modèle équivalent a été calculé avec pour VD l'aire totale de l'espace vocalique. Aucune différence significative n'a pu être mise en évidence.

Malgré cette absence de résultats systématiques, nous remarquons, comme à la Figure 1, que certains locuteurs présentent un espace vocalique plus grand dans la condition allongée (vert) que dans la condition assise (rouge).

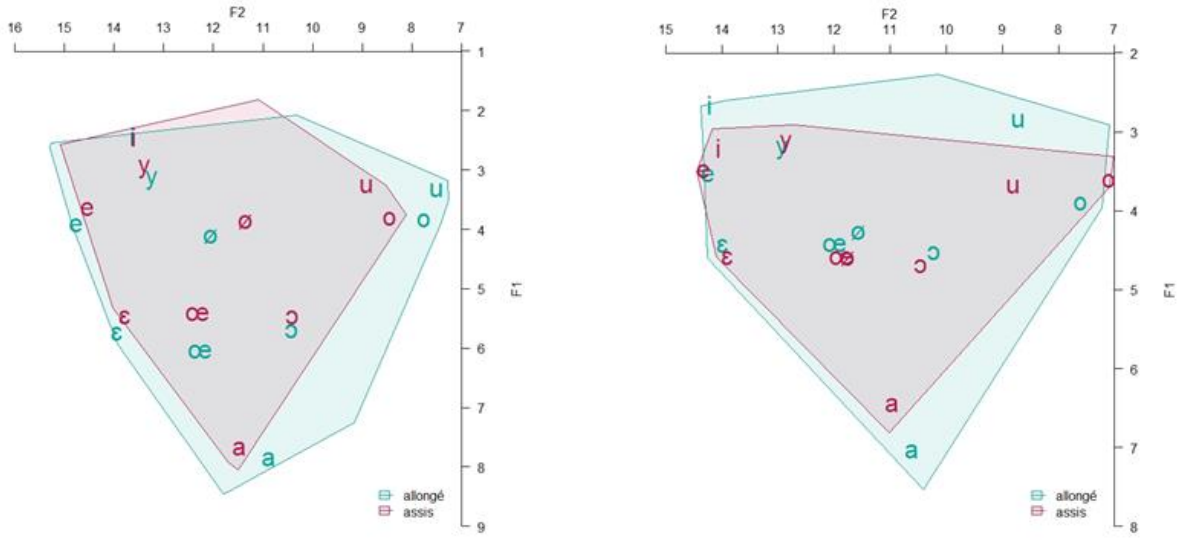



FIGURE 1 : Espace acoustique (Bark) des locuteurs KI (gauche) et SV (droite).

Afin d'évaluer un éventuel effet de la position du sujet sur le F0 moyen et sur l'amplitude de la variation de F0 (écart-type) en demi-tons mesurés sur l'extrait de texte, ces deux variables ont été incluses dans un modèle similaire au précédent. L'analyse ne conclut pas à un effet significatif.

Une analyse impliquant les facteurs Voyelle et Position et la VD durée des voyelles ne fait apparaître aucun effet significatif ni aucune interaction. La comparaison des valeurs de débit en fonction de la position ne renvoie pas non plus de résultats significatifs.

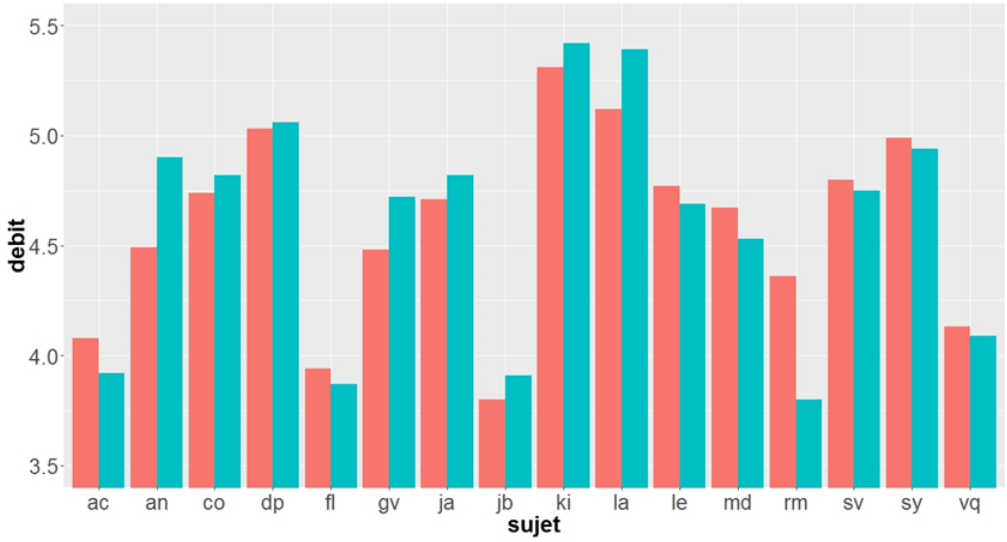



FIGURE 2 : Débit en syllabes par seconde par locuteur et par condition : allongé (rouge) ; assis (vert).

Comme le montre la Figure 2, l'effet de la position corporelle sur le débit de parole agit tantôt dans un sens, tantôt dans l'autre. Il est cependant intéressant de noter que, globalement, le débit propre à chaque locuteur est maintenu.

Enfin, notre dernière variable dépendante, le quotient ouvert (moyen sur la voyelle /a/) ne donne pas non plus lieu à une différence significative en fonction de la position.

# 4 Discussion

D'après Kitamura et al. (2005), les changements que subissent les différents organes de la parole sous l'effet de la force gravitationnelle entraînent des changements dans la configuration du conduit vocal. Ainsi, les paramètres acoustiques de la parole devraient également être modifiés à moins que les locuteurs mettent en place des stratégies visant à rétablir l'intégrité des cibles acoustiques. Les résultats de notre étude suggèrent que les fréquences formantiques des voyelles orales du français ne sont pas significativement impactées par la posture corporelle. Ce constat nous amène à conclure que les locuteurs sont capables, en position allongée, d'ajuster les paramètres articulatoires afin de conserver les aspects acoustiques de la parole.

Les valeurs de F0 restent similaires entre les deux conditions. Ces résultats corroborent les recherches effectuées par Flory et Nolan (2015). Selon ces auteurs, les valeurs de F0 sont plus hautes en position allongée sur le ventre mais restent similaires en position allongée sur le dos et en position assise. Nous retrouvons les mêmes constats pour F1 et F2 : leurs valeurs ne diffèrent pas significativement selon la posture corporelle. Ces résultats sont cohérents avec l'étude de Hoedl (2015). L'impact de la posture corporelle sur la position et la dispersion de l'espace vocalique global varie légèrement d'un locuteur à l'autre. Sur un plan descriptif, il semblerait que la position allongée induise un espace vocalique plus dispersé chez certains. En particulier, contrairement à ce qu'ont montré Kitamura et al. (2005), la position plus basse de la voyelle /a/ chez certains de nos locuteurs (dont ceux de la Figure 1) en condition allongée laisse à penser que l'ouverture de la mâchoire n'a pas été affectée dans notre cas.

Nous avons mené une étude préliminaire sur l'analyse du signal EGG en mesurant le quotient ouvert. Cette première analyse indique qu'il n'y a pas de différence significative entre les deux postures corporelles. Ce résultat reste néanmoins très provisoire puisqu'il conviendrait d'étendre les mesures effectuées aux autres voyelles et éventuellement à l'extrait de texte. Il sera en outre utile d'explorer d'autres options pour caractériser plus finement le signal EGG (Henrich et al., 2004).

La stabilité des paramètres acoustiques de la parole s'explique probablement par le fait que les sujets compensent les modifications articulatoires survenant lors d'un changement de position corporelle. Néanmoins, nous devons insister sur le fait que les participants à cette étude n'avaient pas de trouble particulier. Nous anticipons que les personnes les plus directement concernées par un enregistrement en position allongée – sujets somniloques ou patients soumis à un examen d'imagerie médicale (scanner, IRM, etc.) – pourraient présenter des différences nettement plus marquées. En effet, les ressources nécessaires en termes de tonus musculaire peuvent ne pas être disponibles selon l'état du patient et rendre la compensation articulatoire plus difficile.

# 5 Conclusion

Les résultats présentés dans cette étude contribuent à élargir les connaissances dans le domaine de la phonétique et notamment, de l'impact que peut avoir notre posture corporelle sur la parole. Même si la gravité semble impacter l'articulation, le sujet est doté d'une capacité à contrer les effets de la gravité et ainsi, à conserver les aspects acoustiques de la parole. Un prolongement logique de notre étude consisterait à analyser d'autres positions plus contraignantes (par exemple, la position ventrale) et à inclure d'autres populations. Il serait également pertinent de collecter des données articulatoires.

# Références

ARNULF I., UGUCCIONI G., GAY F., BALDAYROU E., GOLMARD J-L., GAYRAUD F., DEVEVEY A. (2017). What Does the Sleeping Brain Say? Syntax and Semantics of Sleep Talking in Healthy Subjects and in Parasomnia Patients. *Sleep*, 40(11).

ENGWALL O. (2006). Assessing MRI measurements : Effects of sustenation, gravitation and coarticulation. *Psychology Press*, 301- 314.

FLORY Y., NOLAN F. (2015). The influence of body posture on the acoustic speech signal. *Proceedings of the 18th International Congress of Phonetic Sciences*.

HENRICH N., D'ALESSANDRO C., DOVAL B., CASTELLENGO M. (2004). On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. *Journal of the Acoustical Society of America* 115(3), 1321-1332.

HOEDL P. (2015). Defying gravity: formant frequencies of English vowels produced in upright and supine body position. *Proceedings of the 18th International Congress of Phonetic Sciences*.

KITAMURA T., TAKEMOTO H., HONDA K., SHIMADA Y., FUJIMOTO I., SYAKUDO Y., et al. (2005). Difference in vocal tract shape between upright and supine postures: Observations by an open-type MRI scanner. *Acoustical Science and Technology* 26(5), 465- 468.

SHILLER D-M., OSTRY D-J., GRIBBLE P-L. (1999). Effects of Gravitational Load on Jaw Movements in Speech. *Journal of Neuroscience* 19(20), 9073- 9080.

STONE M., STOCK G., BUNIN K., KUMAR K., EPSTEIN M., KAMBHAMETTU C., ET AL. (2007). Comparison of speech production in upright and supine position. *The Journal of the Acoustical Society of America* 122(1), 532- 541.

TIEDE M.K., MASAKI S., VATIKIOTIS-BATESON E. (2000). Contrasts in speech articulation observed in sitting and supine conditions. *Proceedings of the 5th Seminar on Speech Production*, 25- 28.

TRASER L., BURDUMY M., RICHTER B., VICARI M., ECHTERNACH M. (2013). The effect of supine and upright position on vocal tract configurations during singing — A comparative study in professional tenors. *Journal of Voice* 27(2), 141-148.

WRENCH A.A., CLELAND J., SCOBBIE J-M. (2011). An ultrasound protocol for comparing tongue contours: upright vs supine. *Proceedings of the 17th ICPhS Hong Kong,* 2161- 2164.

# Jugements sur le nombre de syllabes et coordination temporelle des gestes articulatoires

Anisia Popescu, Ioana Chitoran
UFR Linguistique et Clillac-ARP, 8 Place Paul-Ricoeur, 75013 Paris, France
anisia.popescu@univ-paris-diderot.fr, ioana.chitoran@univ-paris-diderot.fr

## RESUME

L'article traite de la relation entre la durée acoustique des rimes et la connaissance phonologique des locuteurs concernant sur le nombre de syllabes dans un mot. L'étude présentée est une réplication et une extension d'une étude portant sur cette relation dans le cas des sesquisyllabes de l'Anglais Américain. L'extension est basée sur des résultats portant sur la coordination temporelle des gestes articulatoires des consonnes liquides en coda de syllabe. Le cas des sesquisyllabes nous permet de tester l'hypothèse d'une représentation partagée entre le contrôle moteur et la phonologie. Les résultats sont en conformité avec cette hypothèse et montrent que cette représentation a une variabilité inter-locuteurs.

## ABSTRACT

**Syllable count judgments and temporal organization of articulatory gestures.**

The paper investigates the relationship between the acoustic duration of rimes and speakers' phonological knowledge of syllable count judgments. It is a replication and an extension of an earlier study explaining this relationship in the case of American English sesquisyllables. The extension is based on results pertaining to the temporal coordination of articulatory gestures in liquid coda consonants. The case of sesquisyllables allows us to test the hypothesis that speakers share a common representation for speech motor control and phonological knowledge. Our results are consistent with this hypothesis and show that this representation varies between speakers.

MOTS-CLES : sesquisyllabes, coordination articulatoire, consonnes liquides
KEYWORDS: sesquisyllables, articulatory coordination, liquid consonants

# 1 Introduction

Les locuteurs natifs ont des intuitions robustes sur le nombre de syllabes dans un mot. Lavoie & Cohn 1999 et Tilsen & Cohn 2016 ont montré qu'en anglais américain il y a une classe de mots avec un noyau composé soit d'une voyelle tendue soit d'une diphtongue suivie d'une coda liquide (*feel, fire*) pour laquelle les locuteurs ont des jugements variables sur le nombre


10.21437/JEP.2018-66

de syllabes. Cette classe de mots a été dénommée *sesquisyllabes* (Lavoie & Cohn 1999). Tilsen & Cohn 2016 ont montré qu'il y a une corrélation entre les jugements sur le nombre de syllabes (**JNS**) et la durée acoustique des rimes – la durée acoustique diffère entre des mots associés à des jugements monosyllabiques et des jugements dissyllabiques. Ils proposent sur la base de ces résultats qu'une représentation commune est utilisée pour les processus phonologiques comme les JNS et le contrôle moteur de la parole.

L'étude présentée ici maintient comme hypothèse la représentation partagée entre contrôle moteur et phonologie et propose de poursuivre sa vérification par une réplication et une extension de Tilsen & Cohn 2016. Nous proposons dans cette étude l'extension de l'hypothèse au cas des codas liquides complexes, sur la base des résultats obtenus par Marin & Pouplier 2010 pour l'anglais américain. Ces résultats montrent qu'en anglais américain l'organisation temporelle des gestes articulatoires dans le cas des codas complexes liquides présente un patron de coordination qui est différent de celui des codas complexes sans liquides. En coda non liquide, les gestes articulatoires de la consonne (nasale ou occlusive) adjacente au noyau vocalique ont un patron de coordination *en opposition de phase* à la fois avec la voyelle précédente et avec les consonnes suivantes. Ce n'est pas le cas pour les consonnes liquides. Les liquides en coda présentent un patron de coordination *en phase* avec la voyelle et *en opposition de phase* avec les consonnes adjacentes. Quand la complexité phonotactique augmente (CVC : *feel vs. CVCC : field)* cette différence de patrons de coordination se traduit acoustiquement par un raccourcissement significatif de la durée de la voyelle dans le cas des codas complexes liquides (durée /i/$_{field}$ < durée /i/$_{feel}$). Le raccourcissement est présent mais n'est pas significatif dans le cas des nasales ou des occlusives (durée /i/$_{beams}$≈ durée /i/$_{beam}$). Si la durée de la rime a une influence sur les JNS, le raccourcissement significatif dans le cas des mots avec des codas liquides aura une influence sur les JNS des items avec des codas complexes. Pour tester cette hypothèse nous avons répliqué Tilsen & Cohn 2016 en rajoutant des items avec des codas complexes.

## 1.1 Méthodologie

Nous avons mis en place une expérience dont le design expérimental est identique à celui de Tilsen & Cohn 2016 à quelques modifications près, expliquées ci-dessous. 20 locuteurs natifs de l'anglais américain ont participé à cette expérience, tous étudiants en première année à l'Université de Chicago. L'expérience était composée de deux taches séquentielles : une tache de production suivie d'une tache de jugements sur le nombre de syllabes. Au moment de la partie production, les participants n'étaient pas au courant de la partie suivante, soit celle de jugement. La tache de JNS a été modifiée par rapport au design original : Tilsen & Cohn 2016 utilisent une barre graduée de 1 à 2 pour solliciter les JNS. Pour éviter les complications liées à la conversion de données continues en données catégorielles peut s'avérer compliquée. C'est pourquoi nous avons opté pour un choix parmi trois options de réponse : (1) **1** pour un mot considéré monosyllabique, (2) **1.5** pour un JNS intermédiaire – un mot de plus d'une syllabe mais moins de deux syllabes, et (3) **2** pour un mot considéré dissyllabique. Comme dans Tilsen & Cohn 2016, les instructions justifiaient un choix de 1.5 pour un JNS. Trois des 20 locuteurs ont été éliminés car ils ont donné des jugements incohérents pour les items de contrôle.

Les stimuli cibles étaient composés de triplets *syllabe ouverte (CV) – coda simple (CVC1) – coda complexe (CVC1C2)* avec un noyau vocalique consistant soit d'une voyelle tendue soit d'une diphtongue suivie soit d'une consonne latérale soit d'une consonne rhotique et d'une deuxième consonne : *fee - feel – field, tie – tire – tired*. Dans le cas des codas complexes, la majorité des stimuli étaient dimorphémiques (*10 items mono-morphémiques, 41 items dimorphémiques) : 23 formes du passé – pooled, peaked (12 avec liquides ; 6 avec nasales/occlusives) ; 18 formes du pluriel – meals, stains (6 avec liquides ; 5 avec nasales /occlusives)*. Aux stimuli cible nous avons ajouté trois types de paires de contrôles : (1) voyelle relâchée suivie d'une consonne liquide, nasale ou occlusive (*gull-gulp, bin-bins, peak-peaked)* ; (2) voyelle tendue/diphtongue suivie d'une consonne nasale ou occlusive (*zoo – zoom – zoomed, pea -peak – peaked)* et (3) des mots clairement disyllabiques (*public, bacon*).

Un enregistreur Zoom H4NPRO a été utilisé pour recueillir les données de production. L'analyse acoustique a été faite sous Praat (Boersma & Weenink 2015). Comme les frontières entre voyelle et liquide ne sont pas clairement identifiables, nous avons comparé la durée de la voyelle + la consonne suivante (VC1) pour tous les items. La mesure envisagée pour mesurer le degré de raccourcissement de la durée de la voyelle dans chaque paire coda simple - coda complexe est le ratio $VC1_{complexe}/VC1_{simple}$ ($D[i:ɬ]_{field}$/ $D[i:ɬ]_{feel}$). Les valeurs attendues se trouvent dans l'intervalle [0.5, 1]. Plus la valeur est proche de 1 moins le raccourcissement est important. Pour la tache de jugement une application a été créée, suivant les indications de Tilsen & Cohn 2016. Contrairement à ces derniers, nous n'avons pas écarté les réponses avec des temps de réponse au-delà de 5 secondes n'ont pas été écartées. A la place nous avons enregistré les temps de réponse pour chaque item. Toutes les analyses statistiques ont été faites avec le logiciel R (R Core Team 2013) en utilisant les packages des modèles linéaires mixtes (lme4) et des modèles de régression ordonnés (ordinal).

## 1.2 Prédictions

Pour les mesures de durée on s'attend à des ratio $VC1_{complexe}/VC1_{simple}$ inférieurs à 1 pour tous les types de consonnes en coda, avec un ratio plus éloigné de 1 dans le cas des liquides que dans le cas des nasales et des occlusives ($Ratio_{liquides} < Ratio_{nasales/occlusives} < 1$).

Basé sur Tilsen & Cohn 2016, avec le rajout d'une deuxième consonne en coda, on s'attend à une augmentation des JNS entre les items CVC et CVCC indépendamment du type de consonne (liquide, nasale, occlusive). La différence de JNS observée entre les codas simples liquides d'une part et les codas simples nasales ou occlusive d'autre part sera préservée dans le cas des codas complexes.

Par contre, si on tient compte des résultats de Marin & Pouplier 2010, le rajout de la deuxième consonne va déclencher un raccourcissement significatif de la voyelle dans le cas des liquides mais pas dans le cas des nasales ou occlusives. Ceci résulterait dans une durée raccourcie de la rime dans le cas des codas complexes liquides. On s'attend alors à ce que les locuteurs augmentent les JNS dans le cas des nasales ou occlusives, mais pas dans le cas des liquides. Ces prédictions sont illustrées dans le Tableau 1.

Un choix de 1.5 pour un JNS pourrait s'avérer contre-intuitif pour les participants. Même si les locuteurs n'optent pas tous pour un jugement intermédiaire (*JNS=1.5*) dans le cas des codas liquides, le temps de réponse pourrait indiquer une différence de représentation entre les liquides et les nasales ou occlusives : on s'attend à des temps de réponse plus longs dans le cas des liquides.

A part la durée et le temps de réponse nous avons considéré trois autres paramètres qui pourraient avoir une influence sur le choix de JNS : la fréquence lexicale des items (d'après le Corpus COCA, Davies 2008), la composition morphémique (*mono- ou di-morphémique*), le type du deuxième morphème (formes du passé ou du pluriel) et l'orthographe (*le nombre de voyelles écrites non-consécutives: field – 1V ; pooled – 2V*).

| | *JNS attribué : 1 simple* | *JNS prédit : complexe* | *JNS attribué : 1.5 simple* | *JNS prédit : complexe* | *JNS attribué : 2 simple* | *JNS prédit : complexe* |
|---|---|---|---|---|---|---|
| Coda liquide | **1σ** feel | **1σ** field | **1.5σ** feel | **1.5σ** field | **2σ** - | - |
| Coda nasale | **1σ** pain | **1.5σ** pained | 1.5σ - | - | **2σ** - | - |
| Coda occl. | **1σ** peak | **1.5σ** peaked | 1.5σ - | - | **2σ** - | - |

Tableau 1 : Prédictions des JNS pour les codas complexes en fonction des JNS attribués aux codas simples correspondants (basé sur Tilsen & Cohn 2016 et Marin & Pouplier 2010)

# 2 Résultats

Nous allons présenter les résultats en trois étapes. En 2.1 nous allons confirmer les résultats de Marin & Pouplier 2010 portant sur le raccourcissement acoustique de la voyelle devant les codas liquides. La section 2.2 portera sur les jugements sur le nombre de syllabes (JNS). Finalement, en 2.3 nous allons tester la corrélation entre les deux types de résultats.

## 2.1 Durée acoustique de la voyelle

Une comparaison de modèles linéaires mixtes avec *Consonne_Coda* comme facteur fixe et *Locuteur, Item, Fréquence* comme facteurs aléatoires montre qu'il y a un effet global du type de consonne coda sur le ratio $VC1_{complexe}/VC1_{simple}$ *(p= .025)*. Les séquences V + latérale et V + rhotique ont une durée acoustique significativement plus courte dans des items CVC1C2 que dans des items CVC1 (*Figure 1 gauche*). Les paires avec des séquences V + nasale ou V + occlusive ne montrent pas cette différence (*p=0*.17). Il n'y a pas de différence entre les latérales et les rhotiques *(p= .57)*. Ces résultats sont en accord avec Marin & Pouplier 2010. Une analyse par locuteurs montre que le degré de raccourcissement est variable entre locuteurs. Certains locuteurs ne présentent pas de raccourcissement significatif dans le cas des liquides. Cette variabilité inter-locuteurs devrait se retrouver dans l'attribution des JNS.

Une même analyse sur la durée totale des rimes (ratio $VC1C2_{complexe}/VC1_{simple}$) indique qu'il y a une différence significative (*p= .031*) entre l'allongement de la durée de la rime pour les items avec des séquences V + liquide d'une part, et les items avec des séquences V + nasale ou V + occlusive d'autre part (*Figure 1 droite*). L'allongement des items avec des consonnes nasales ou occlusives est plus grand. Ceci n'est pas surprenant si on prend en compte le raccourcissement significatif de $VC1_{complexe}$ dans le cas des liquides.

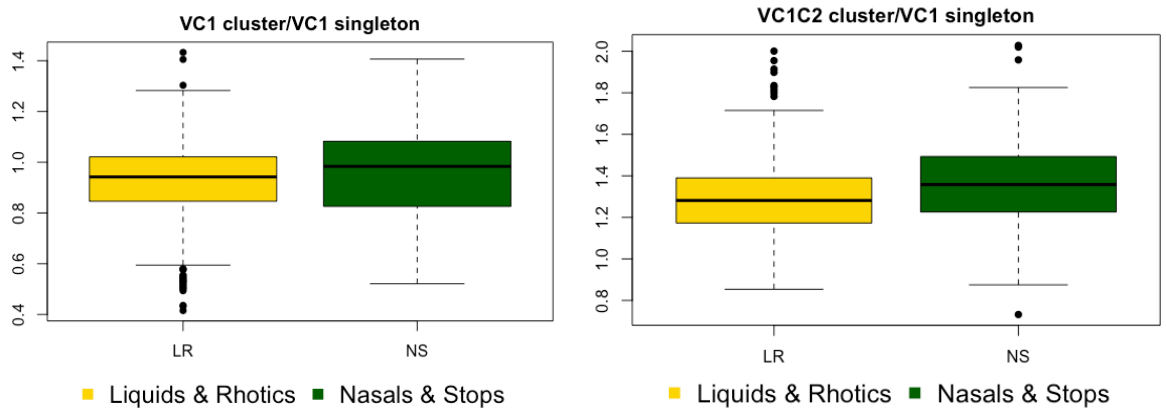



Figure 1 : $VC1_{complexe}/VC1_{simple}$ (gauche) et $VC1C2_{complexe}/VC1_{simple}$ (droite) en fonction du type de consonne en coda (liquide ou nasale/occlusive)

## 2.2 Jugements sur le nombre de syllabes (JNS)

Les résultats des jugements des locuteurs sont plus variables que ceux des données de production. On peut pourtant faire des généralisations. La figure 2 montre le nombre total des trois types de JNS (*1, 1.5 ou 2 syllabes*) attribué aux différents degrés de complexité phonotactique (*coda simple vs. coda complexe*) et aux types de consonne en coda (*latérale, nasale, rhotique ou occlusive*). On peut observer que dans le cas des codas simples des JNS supérieurs à 1 (*JNS=1.5, JNS=2*) sont attribués exclusivement aux items avec une coda liquide. Ces JNS sont attribués par un même groupe de locuteurs qui sont cohérents dans leurs réponses. Ce résultat confirme la variabilité des JNS rapportée pour les sesquisyllabes (Lavoie & Cohn 1999, Tilsen & Cohn 2016).

La figure 3 montre les temps de réponse pour les JNS en fonction de la complexité phonotactique et du type de consonne en coda. On observe que le temps de réponse est plus long pour les codas complexes (*boites de gauche pour chacune des quatre couleurs*) que pour les codas simples (*boites de droite pour chacune des quatre couleurs*). Le temps de réponse pour les codas liquides est plus long que pour les codas nasales et occlusives, dans le cas codas simples aussi bien que des codas complexes. Une comparaison de modèles linéaires mixtes montre qu'il y a un effet du type de consonne en coda sur le temps de réponse: pour les nasales et les occlusives le temps de réponse est plus court (*p=.02, p=.03 respectivement*), mais il n'y a aucune différence entre les latérales et les rhotiques (*p= 0.22*).

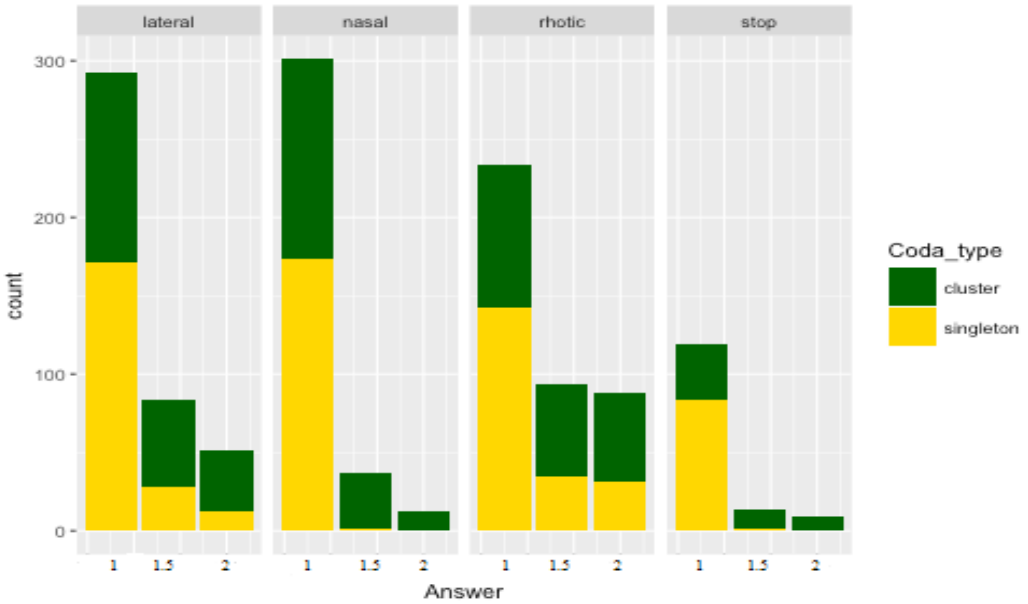



FIGURE 2: Nombre total de JNS (*1, 1.5 resp. 2*) attribué aux différents types de coda (*simple ou complexe)* et types de consonne coda (*latérale, nasale, rhotique ou occlusive*)

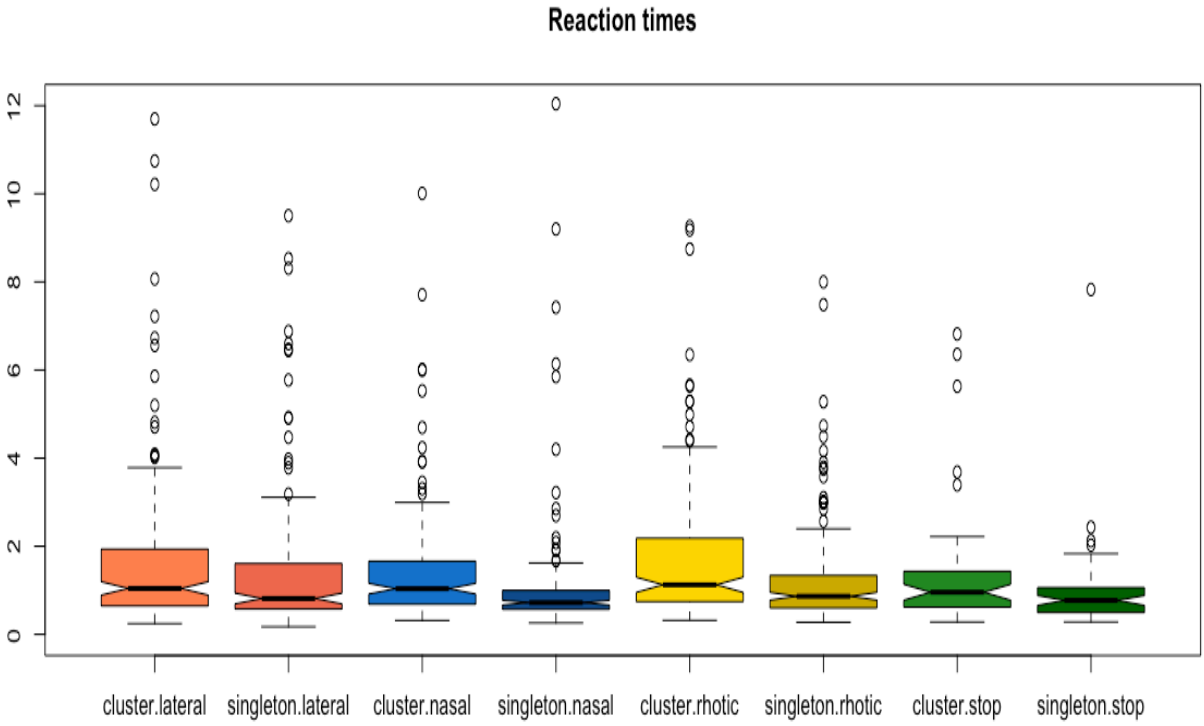



FIGURE 3: Temps de réponse (*en secondes*) pour les JNS pour différentes consonnes coda (*latérale, nasale, rhotique ou occlusive)* et type de coda (*simple vs. complexe*)

Même si les locuteurs n'optent pas tous pour un jugement intermédiaire (*JNS=1.5*) dans le cas des codas liquides, les temps de réponse plus longs montrent que ces mots constituent une classe à part. On observe une différence entre les sesquisyllabes et leurs homologues avec des codas nasales ou occlusives.

Nous avons considéré la possibilité que la fréquence peut aussi influencer l'attribution des JNS. Nous avons trouvé que de manière générale le même groupe restreint de locuteurs attribue des jugements supérieurs à 1.5 aux sesquisyllabes. Une analyse des logarithmes des fréquences à partir du corpus COCA (Davies 2008- ) montre que ce groupe devient plus grand lorsque le mot jugé est moins fréquent.

Dans le cas des codas complexes, la majorité des stimuli sont dimorphémiques. Un test $\chi^2$ montre qu'il y a une différence entre les items monomorphémiques et dimorphémiques *($\chi^2$ =44.711, df = 4, p = 4.566e-09*). Les locuteurs attribuent plus de jugements supérieurs à 1 aux items avec des codas complexes formées de deux morphèmes. Pour le type de morphème on trouve que les locuteurs donnent plus de jugements supérieurs à 1 si le deuxième morphème est une marque du passé (*–ed)* que si c'est une marque du pluriel (*-s) (X-squared = 13.886, df = 2, p = .00096)*. Une cause possible pour la différence entre les deux types de morphèmes pourrait venir de l'orthographe, i.e. la présence ou l'absence d'un graphème en plus (*peeke**d** vs. peaks)*. Nous avons donc considéré le nombre de voyelles non consécutives comme paramètre. Un test $\chi^2$ montre que le graphème supplémentaire a une influence sur les jugements attribués aux items avec des codas complexes, indépendamment du type de consonne en coda. Des modèles cumulatifs mixtes confirment que le nombre et le type de morphèmes ont une influence sur les jugements des locuteurs. Ces résultats ont déterminé le choix d'inclure l'orthographe et la structure morphologique comme facteurs aléatoires dans les modèles cumulatifs présentés en 2.3.

Les résultats portant sur les jugements sur le nombre de syllabe sont donc influencés par la fréquence lexicale, le nombre des morphèmes et l'orthographe. Cependant, les résultats portant sur l'influence de la consonne en coda et les temps de réponse pour chaque JNS montrent que les locuteurs ont une représentation particulière pour la classe des items avec des codas liquides. Dans la section suivante nous allons présenter les résultats sur la corrélation entre la durée acoustique des rimes et le JNS.

## 2.3 Corrélation durée acoustique et JNS

Pour analyser la corrélation entre la durée acoustique et les JNS nous avons utilisé un modèle de régression ordonné cumulatif (*package ordinal*) avec des facteurs aléatoires (*Type de consonne coda, Locuteur, Complexité phonotactique)*. Les résultats montrent que la durée totale des rimes est un bon prédicteur pour les jugements sur le nombre de syllabes (*p<2e-16*). Des JNS supérieurs à 1 sont attribués aux items avec des rimes plus longues (indépendamment du type de coda, simple ou complexe). Pour toutes les données (CV, CVC, CVCC) ensemble les jugements monosyllabiques (JNS=1) sont attribués aux items avec les rimes les plus courtes. Par rapport à la complexité phonotactique, les JNS varient entre locuteurs. Nous observons deux stratégies différentes. Certains locuteurs attribuent les mêmes JNS aux items par paires CVC - CVCC, d'autres augmentent la valeur du JNS de 1 syllabe/1.5 syllabes pour un item CVC à 1.5 syllabes/2 syllabes pour sa contrepartie complexe CVCC. La différence entre les deux groupes est le degré de raccourcissement des séquences VC1 : dans le cas où le JNS reste inchangé entre CVC et CVCC, il y a un degré

plus fort de raccourcissement de VC1 ; dans le cas où le JNS augmente entre CVC et CVCC le raccourcissement est moins marqué. Les deux stratégies sont illustrées dans le tableau 2. Il n'y a donc pas de stratégie universelle d'attribution des JNS, mais à l'intérieur d'une stratégie précise les prédictions sont confirmées.

| | *JNS attribué pour CVC* | *JNS attribué pour CVCC* | *Degré de raccourcissement de la voyelle* |
|---|---|---|---|
| STRATEGIE 1 | 1 (ou 1.5) | 1 (ou 1.5) | haut |
| STRATEGIE 2 | 1 (ou 1.5) | 1.5 (ou 2) | bas |

Tableau 2 : Illustration des deux stratégies utilisée pour l'attribution des JNS en fonction du degré de raccourcissement de la voyelle

Il y a donc une corrélation entre les jugements et la production. La durée totale de la rime est un paramètre qui est pris en compte lors de l'attribution des jugements. Le fait que les deux taches expérimentales sont indépendantes soutient l'hypothèse d'une représentation commune du contrôle moteur de la parole et des jugements phonologiques.

# 3 Conclusion

Dans cette étude nous avons montré une corrélation entre la durée des rimes et les intuitions des locuteurs sur les jugements sur le nombre de syllabe dans un mot. Nous avons répliqué les résultats de Tilsen & Cohn 2016 et nous avons trouve des résultats similaires dans l'extension de la tache aux codas simples vs. complexes. Nous avons montré qu'il y a une grande variabilité inter-locuteurs dans l'attribution des jugements sur le nombre de syllabes. Pourtant, cette variabilité est cohérente avec la variabilité de production. Nos résultats vont donc dans le même sens que ceux de Tilsen & Cohn 2016 et restent consistants avec l'hypothèse d'une représentation partagée entre les processus gérant le contrôle moteur et la phonologie.

Malgré l'accord entre les résultats des deux études on ne peut pas conclure définitivement sur le statut de cette représentation. Nous avons vu, par exemple, que le rôle de l'orthographe ne peut pas être ignoré dans la tache de jugement. En même temps, le paramètre acoustique de la durée de la rime reste une mesure indirecte de la coordination temporelle. Nous envisageons pour continuer, une étude articulatoire qui nous permettrait de tester de manière plus directe le lien entre la coordination temporelle en production et les jugements phonologiques.

# Références


BATES D., MAECHLER, M., BOLKER, B., WALKER, S., (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1-48.

BOERSMA P., WEENINK, D., (2018). Praat : doing phonetics by computer [Computer program]. Version 6.0.08, retrieved in 2015 form http://www.praat.org/

CHISTENSEN R., H., B., (2015). ordinal : Regression Models for Ordinal Data. R package, http://www.cran.r-project.org/package=ordinal/

DAVIES M., (2008- ). The Corpus of Contemporary American English (COCA) : 560 million words. Available online at http://corpus/byu.edu/coca

LAVOIE L.,M., COHN, A., (1999). Sesquisyllables of English : the Structure of vowel-liquid syllables. *In proceedings of the XIVth International Congress of Phonetic Sciences* 109-112

MARIN S., POUPLIER, M., (2010). Temporal Organization of Complex Onsets and Codas in American English : Testing the Predictions of a Gestural Coupling Model. *Motor Control,* 14, 380-407.

R CORE TEAM, (2013). R : a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

TILSEN S., COHN, A., (2016). Shared representations underlie metaphonological judgments and speech motor control, *Laboratory Phonology : Journal of the Association for Laboratory Phonology,* 7(1) : 14, 1-13.

# La voyelle inaccentuée <e> en position initiale : analyses acoustiques et enjeux pédagogiques pour l'anglais L2.

Anne Tortel & Sophie Herment
Aix-Marseille Université, Laboratoire Parole et Langage, CNRS-UMR7059,
5 avenue Pasteur, 13100 Aix en Provence, France
anne.tortel@univ-amu.fr ; sophie.herment@univ-amu.fr

## RESUME

Les travaux sur la réduction vocalique sont nombreux mais les études du <e> inaccentué à l'initial de mot sont quasi inexistantes, notamment lorsqu'il s'agit d'étudier des productions d'apprenants francophones. L'étude de cette voyelle est donc ici réalisée sur un corpus d'anglais oral lu par 10 natifs anglophones et 20 apprenants francophones de l'anglais en analysant les valeurs de F1-F2 et de durées. Les résultats révèlent un changement phonologique en cours intéressant pour l'anglais RP contemporain et les données obtenues avec nos trois groupes de locuteurs permettent de comprendre la distribution et la réalisation de la voyelle inaccentuée <e> par les natifs et les apprenants de l'anglais. Cette étude nous permet également d'établir des perspectives pédagogiques pour l'apprentissage et l'enseignement de l'anglais L2.

## ABSTRACT

**Unstressed vowel <e> in initial position: acoustic analyses and pedagogical issues for ESL.**
Vocalic reduction has been studied in various phonetic and phonological perspectives but the realisation of reduced <e> in initial syllables in English has hardly been dealt with. In this study the realisations of this unstressed vowel produced by 10 native English speakers and 20 French learners of English were analysed using an automatic extraction of F1-F2 values and durations in a corpus of read speech. The results are enlightening as far as phonological change is concerned and the data obtained with the three groups of speakers contribute to understanding the distribution and realisation of unstressed vowels by native speakers and learners of English. This study finally opens on pedagogical perspectives for both learning and teaching ESL.

MOTS-CLES : réduction vocalique, rythme, durée vocalique, apprenants, anglais L2.
KEYWORDS: vocalic reduction, rhythm, vocalic duration, learners, L2 English.

## 1. Introduction

L'acquisition du rythme de l'anglais représente une des difficultés majeures de l'apprentissage de cette langue pour des francophones et tout particulièrement la production des voyelles inaccentuées. A un niveau avancé, les apprenants francophones de l'anglais réalisent souvent très bien les voyelles accentuées, mais pèchent à prononcer correctement les voyelles inaccentuées, surtout lorsqu'elles sont réduites (Tortel, 2009). Cela vient du fait que les deux langues présentent une organisation rythmique complètement différente (Wenk & Wioland, 1982). En anglais l'alternance de temps forts et de temps faibles a pour conséquence sur le système vocalique la réduction de certaines


10.21437/JEP.2018-67

voyelles (Bolinger, 1981). Cette caractéristique semble typique de l'anglais oral et non du français (même si les réalisations des voyelles françaises varient sur le plan phonétique, voir Gendrot & Adda-Decker, 2007). Alors qu'il existe des règles régissant le processus de réduction en anglais, cela reste très difficile pour un apprenant de savoir quelle voyelle il doit réaliser lorsqu'il produit une réduction. En effet, la voyelle d'une syllabe inaccentuée peut avoir sa forme pleine ou sa forme réduite, et lorsqu'elle est réduite, plusieurs réalisations sont parfois possibles. Tel est le cas, par exemple, du mot *enjoy* pour lequel il est possible de prononcer [i, e, ə] (Wells, 2008). Dans une étude fondée sur les données des dictionnaires de prononciation anglaise *Longman Pronunciation Dictionary* (Wells, 2008) et *English Pronunciation Dictionary* (Jones, 2011), désormais LPD3 et EPD17, Herment (2010) montre que des indices permettent de prédire les réalisations des voyelles réduites en positions médiane et finale, mais que très peu de tendances se dégagent pour la position initiale, et particulièrement en ce qui concerne le graphème <e>. De nombreuses combinaisons de variantes apparaissent dans les deux dictionnaires, qui ne sont pas toujours en accord : pour ne prendre qu'un exemple, la première syllabe du mot *emotion* se prononce [i] ou [ə] selon LPD3 et [ɪ] selon EPD18 (*cf.* tableau 1). Nous proposons donc de nous intéresser dans un premier temps à la réalisation de la voyelle <e> en position initiale inaccentuée dans un corpus de parole lue par des locuteurs natifs, afin de comparer nos résultats aux données dictionnairiques et de tenter d'établir des tendances robustes pour la réalisation de cette voyelle. Dans un second temps, nous examinons les réalisations de cette voyelle par deux groupes d'apprenants francophones sur le même corpus lu afin de les comparer aux productions des natifs et de montrer quelles sont les enjeux pédagogiques que l'on peut en retirer.

# 2. Corpus

## 2.1 AixOx

Les recherches présentées ici ont été réalisées à partir du corpus multilingue AixOx (Herment *et al*., 2012, 2014). Ce corpus fait partie de la base de données OMProDat (Open Multilingual Prosodic Database, Hirst *et al*., 2013), qui constitue une collection d'enregistrements suivant le protocole Eurom 1 (Chan *et al*., 1995) dans plusieurs langues. AixOx est composé de productions orales enregistrées en chambre sourde par des natifs anglophones et francophones, des apprenants anglophones du français et des apprenants francophones de l'anglais. 40 textes composés de 5 phrases chacun et racontant des histoires du quotidien ont été lues par les locuteurs. Pour les besoins de cette étude, seules les productions des natifs anglophones et des apprenants francophones de l'anglais ont été analysées. Cela représente plus de 20 heures de parole lue pour 30 locuteurs : 10 locuteurs natifs (5 hommes, 5 femmes) âgés de 18 à 26 ans, tous originaires d'Oxford et y résidant, et parlant une variété d'anglais qualifiée de « RP contemporain » selon les experts de la *British Library*[1] ; 10 apprenants francophones de l'anglais (5 femmes, 5 hommes) de niveau C (expérimenté) selon le CECRL[2], âgés de 18 à 32 ans, originaires de la région PACA ; 10 apprenants francophones de l'anglais de niveau B (indépendant) selon le CECRL, âgés de 18 à 32 ans, 8 d'entre eux étant de la région PACA et 2 originaires de la région parisienne. 36 mots du corpus contiennent la voyelle <e> à l'initiale de syllabe. Parmi eux, 3 apparaissent à plusieurs reprises soit dans le corps

[1] http://www.bl.uk/learning/langlit/sounds/find-out-more/received-pronunciation/, consulté le 25 janvier 2018. RP signifie « Received Pronunciation ». Il s'agit de la variété britannique standard.

[2] Cadre Européen Commun de Référence pour les Langues, Unité des politiques linguistiques de Strasbourg, www.coe.int/lang-CECR, consulté le 5 février 2018

d'un même texte, comme pour *department (2 fois)* et *delivery (2 fois)*, soit dans des textes différents, comme pour *department* (apparaissant 6 fois au total) ou encore *because* (2 fois). Cela donne un total de 46 occurrences par locuteur.

### 2.2 Données dictionnairiques et hypothèse

Pour les 36 mots du corpus, les prononciations du <e> présent dans la syllabe initiale ont été relevées dans les deux dictionnaires mentionnés plus haut (LPD3 & EPD18). Le tableau 1 en montre un extrait pour les premiers mots. On note une grande variabilité, et le manque de consensus entre les deux ouvrages est frappant. La possibilité est donnée par LPD3 de prononcer un schwa [ə] pour toutes les occurrences, ce qui n'est pas le cas de EPD18. D'un point de vue diachronique, la confrontation avec les données de 1967 (Jones, 1967[3]) est parlante : pour la plupart des mots étudiés, [ə] n'est pas donné comme une variante possible. Nous formulons donc l'hypothèse que les locuteurs de RP contemporain montrent une tendance à la centralisation des voyelles initiales inaccentuées et à prononcer [ə].

| Words | LPD3 | EPD18 | Jones 1967 |
|---|---|---|---|
| Because | [i, ə] | [ɪ, ə] | [i, ə] |
| Behind | [i, ə] | [ɪ, ə] | [i] |
| Believe | [i, ə] | [ɪ, ə] | [i, (ə)] |
| Behave | [i, ə] | [ɪ, ə] | [i, (ə)] |
| Department | [i, ə] | [ɪ, ə] | [i] |
| Delivered | [i, ə] | [ɪ, ə] | [i] |
| Delivery | [i, ə] | [ɪ, ə] | [i] |
| December | [i, ə] | [ɪ, ə] | [i, i:] |
| Devoured | [i, ə] | [ɪ, ə] | [i] |
| Destroyed | [i, ə] | [ɪ, ə] | [i] |
| Dependent | [i, ə] | [ɪ, ə] | [i] |
| Effect | [ə, i] | [ɪ] | [i] |
| Electric | [i, ə] | [i] | [i, (ə)] |
| Emergency | [i, ə] | [ɪ, i:] | [i] |
| Emotion | [i, ə] | [ɪ] | [i] |
| Enjoyed | [i, e, ə] | [ɪ, ə] | [i, (e)] |

TABLEAU 1 : exemple de prononciation d'un échantillon de mots contenant la voyelle <e> inaccentuée à l'initiale de mot, selon les 3 dictionnaires de référence

## 3. Méthode

Afin de tester l'hypothèse ci-dessus, et de voir ce qu'il en est des productions des apprenants, nous avons effectué des mesures acoustiques automatiques sur les voyelles étudiées. Les 46 occurrences du corpus ont été mesurées chez nos 30 locuteurs (10 natifs, 10 apprenants indépendants et 10 apprenants expérimentés, voir 2.1) pour un total de 1325 voyelles après correction : quelques

[3] Le plus ancien dictionnaire de prononciation disponible dans notre université

occurrences ont dû être écartées car non réalisées comme *December* prononcé [tˈsembə] ou erronées (à cause d'une qualité de voix particulière par exemple). Le corpus a été phonétisé et aligné automatiquement avec SPPAS (Bigi, 2012), qui génère des fichiers TextGrid, puis corrigé manuellement sous PRAAT (Boersma, Weenink, 2001). Les mesures de 4 paramètres ont ensuite été extraites pour chaque voyelle à l'aide d'un script PRAAT : F1, F2, F3 et durée. Les valeurs correspondent à la moyenne sur la durée de la voyelle. Les mesures ont ensuite été traitées dans NORM (Thomas, Kendall, 2007). Afin d'éviter les différences inter-locuteur résultant de la variation de longueur du conduit vocal, la procédure de normalisation Lovanov (1971) a été appliquée pour faciliter la comparabilité des données.

# 4. Résultats

## 4.1 Analyses formantiques

### *4.1.1 Locuteurs natifs*

Les résultats obtenus pour les valeurs formantiques chez les natifs sont montrés dans la figure 1. Les valeurs de F1 et F2 (en Hz) sont données en ordonnée et en abscisse respectivement, ce qui permet d'obtenir la représentation des voyelles sur un diagramme représentant l'espace vocalique (tronqué car les valeurs de F1 et F2 ont été ajustées pour plus de clarté sur la figure, ainsi tout l'espace vocalique n'est pas représenté ici). La valeur de F1 est corrélée avec le degré d'ouverture des voyelles (les valeurs basses correspondent à un faible degré d'aperture) et la valeur de F2 correspond à la position de la langue (réalisations antérieures vers la gauche, postérieures vers la droite). Les valeurs de F3 servent à la procédure de normalisation. On voit clairement sur la figure 1 que la plupart des voyelles examinées sont concentrées vers la zone centrale haute, allant de 320 à 420 Hz pour F1 et de 1300 à 1700 Hz pour F2.

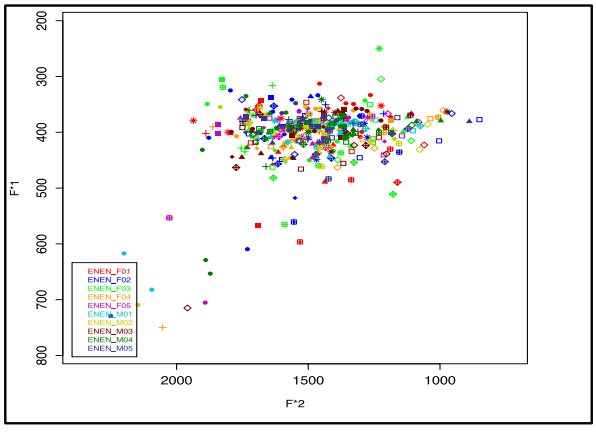



FIGURE 1: représentation dans NORM des valeurs normalisées de F1 et F2 pour les 413 voyelles étudiées et produites par les natifs

### *4.1.2 Apprenants*

Les mêmes analyses formantiques ont été effectuées pour chacun des groupes d'apprenants (B et C) afin de les comparer à celles des natifs. Le but est d'obtenir une contribution pédagogique pour l'enseignement et l'apprentissage de l'anglais quant à la distribution et la production de cette voyelle. La figure 2 montre les résultats pour le groupe d'apprenants de niveau C. On constate des

réalisations vocaliques très proches de celles des natifs anglais avec la production d'une voyelle inaccentuée centrale et haute, sensiblement plus haute que celle des natifs. La concentration des voyelles inaccentuées est légèrement plus éparse avec des valeurs allant de 300 à 400 Hz pour le F1 et de 1300 à 1700 Hz pour le F2. On note un nombre plus important de valeurs dissidentes (que nous nommerons par le terme anglais « outliers »).
Les résultats (figure 3) pour le groupe d'apprenants de niveau B diffèrent de façon remarquable. La figure montre des voyelles bien plus éparpillées avec des valeurs formantiques allant de 320 à 520 Hz pour le F1 et de 1400 à 2000 Hz pour le F2. Les voyelles sont plus antérieures et plus basses. Cette figure montre que ce groupe d'apprenants réduit beaucoup moins la voyelle inaccentuée <e> et ce pour la plupart des occurrences.

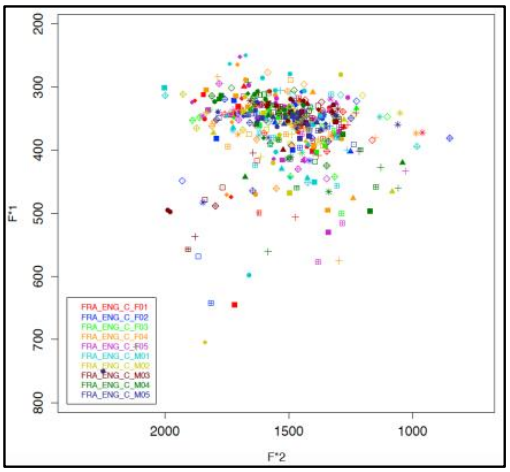



FIGURE 2 : valeurs normalisées F1 et F2, apprenants niveau C

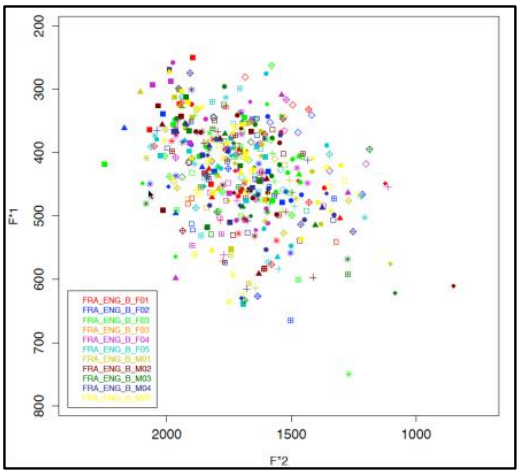



FIGURE 3 : valeurs normalisées F1 et F2, apprenants niveau B

Lorsque l'on compare les résultats des figures 1, 2 et 3, on observe une gradation des valeurs formantiques montrant ainsi les différences de niveau entre les deux groupes d'apprenants mais aussi entre les apprenants et le groupe des natifs.

### *4.1.3 Quelle tendance vocalique pour quel groupe ?*

Nos valeurs étant normalisées, il s'avère difficile de les comparer avec les valeurs que l'on trouve dans la littérature. Nous avons donc pris les valeurs brutes et calculé les moyennes en séparant hommes et femmes : le tableau 2 ci-dessous donne les valeurs moyennes brutes de F1 et F2 pour /ɪ/, /e/, /ʊ/ et /ɜ:/ en discours suivi selon Deterding (1997), pour /i/ selon Herment, Turcsan (soumis), ainsi que les valeurs relevées dans notre corpus (deux dernières lignes).
Ces moyennes montrent que la voyelle type réalisée par nos natifs semble être une voyelle très haute, même légèrement plus haute que /ɪ/. Il s'agit d'une voyelle centrale, plus centrale que /i/, et de façon très intéressante, on note que la valeur de F2 est très proche de celle donnée pour /e/, cette dernière étant pourtant beaucoup plus basse. Cela veut dire que nous avons une voyelle centrale qui tend vers l'avant plutôt que vers l'arrière de l'espace vocalique. Il est difficile de dire s'il s'agit d'un schwa puisqu'il est connu que la qualité de schwa varie de façon conséquente selon les contextes (Browman, Goldstein, 1992, Bates, 1995, Flemming, 2009, Herry-Bénit *et al.*, 2009, Heselwood, 2007, entre autres). Nos résultats confirment donc l'hypothèse de départ d'une centralisation en anglais RP contemporain de la voyelle <e> en position initiale inaccentuée.

Les voyelles des apprenants du groupe C ont des valeurs proches de celles du [i], ce qui n'est pas vraiment surprenant puisque cette voyelle est proche de la voyelle française /i/. Les valeurs pour les apprenants de niveau B ne sont pas données car la déviation standard est trop importante, la moyenne n'est donc pas pertinente pour l'analyse.

| | Hommes | | Femmes | |
|---|---|---|---|---|
| | F1 | F2 | F1 | F2 |
| /ɪ/ | 367 | 1757 | 384 | 2174 |
| /e/ | 494 | 1650 | 719 | 2063 |
| /ʊ/ | 379 | 1173 | 410 | 1340 |
| /ɜː/ | 478 | 1436 | 606 | 1695 |
| [i] | 393 | 1962 | 406 | 2224 |
| **Natifs** | **300** | **1634** | **374** | **1930** |
| **Apprenants C** | **438** | **1816** | **453** | **2111** |

TABLEAU 2 : valeurs moyennes brutes des voyelles hautes telles que trouvées dans la littérature et relevées dans notre corpus.

## 4.2 Durées vocaliques

La qualité de la voyelle réalisée est importante dans la perception du rythme en anglais, mais la quantité l'est également. Peterson, Barley (1952) expliquent qu'en anglais, en périphérie, en contexte inaccentué et à débit rapide, les voyelles ont tendance à être réduites en terme de durées car elles subissent une forte réduction vocalique. Nous pouvons donc prédire que, si les réductions vocaliques sont mal produites ou absentes chez les apprenants, les durées vocaliques des groupes d'apprenants seront plus importantes que celles des natifs. Les mesures de durées ont donc été relevées pour les voyelles examinées.

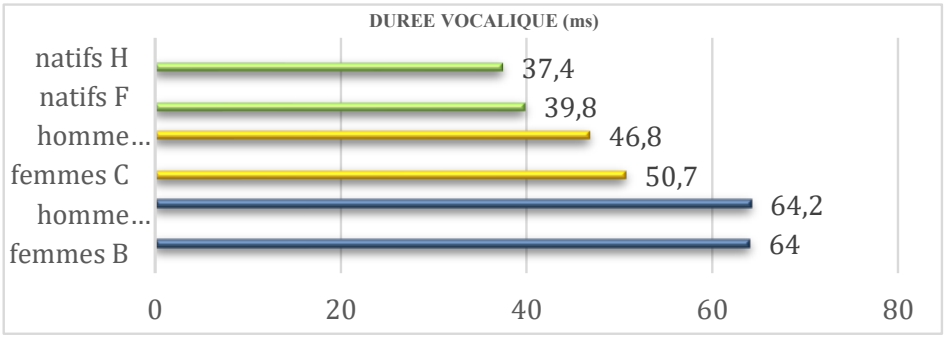



FIGURE 4: représentation des durées vocaliques des réalisations du corpus

La figure 4 regroupe la moyenne en millisecondes des productions pour chacun des groupes natifs, apprenants C et apprenants B, hommes et femmes. On constate une nette gradation avec une augmentation des valeurs allant d'une moyenne de 38,6 ms pour les natifs, 48,75 ms pour les apprenants C, et jusqu'à 64,1 ms pour le groupe des apprenants B. Les productions du groupe d'apprenants indépendants (B) sont plus d'une fois et demi plus longues que celles des natifs. Les durées relevées pour les apprenants C montrent un léger allongement de la voyelle réduite. Cela corrobore les résultats trouvés lors des analyses formantiques, où nous avons expliqué que la figure 3 montre que les apprenants niveau B font beaucoup moins de réductions que les natifs et les apprenants C. Cela est probablement dû à la réalisation d'une voyelle française au lieu d'une réduction vocalique. La durée est donc un élément pertinent qui permet de distinguer les 3 différents groupes. Il est également intéressant de constater qu'en moyenne, les femmes font des voyelles

légèrement plus longues que les hommes chez les natifs comme chez les apprenants C, ce qui va dans le sens de plusieurs études qui montrent que les durées vocaliques sont plus longues chez les femmes que les hommes dans plusieurs langues comme l'allemand, l'anglais ou le suédois (Hillenbrand *et al*., 1995, Simpson, 1998, Ericsdotter, Ericsson, 2001). Si nos apprenants de niveau B calquent des voyelles françaises sur les voyelles censées être réduites en anglais, ceci expliquerait alors pourquoi pour ce groupe la durée est quasiment la même entre hommes et femmes.

# 5. Discussion et perspectives

Les résultats obtenus permettent de dégager des éléments enrichissants concernant le processus allophonique. Les données formantiques sur l'analyse de la voyelle inaccentuée <e> montrent que la prononciation tend vers une voyelle réduite centrale et haute en position initiale de mot et amènent à une neutralisation de l'opposition des voyelles /ə/, /ɪ/ et /e/. Le fait que la voyelle schwa agisse comme une sorte de voyelle centrale variant en fonction du contexte n'est pas nouveau. En revanche, les résultats sur la prononciation de la voyelle inaccentuée <e> à l'initiale de mot apportés par les données dictionnairiques montrent un changement phonologique en cours : la tendance à la centralisation des voyelles réduites en anglais RP contemporain. Ils apportent aussi de nouveaux éléments pédagogiques quant à l'apprentissage de l'anglais : la possibilité de prononcer un /ə/ pour tous les mots contenant la voyelle inaccentuée <e> à l'initiale, ce qui va permettre de faciliter les choses pour les apprenants, mais aussi pour les enseignants, qui se trouvent souvent bien embarrassés pour enseigner le phénomène complexe de la réduction ou non des voyelles inaccentuées en position initiale. L'analyse des durées vocaliques permet de mettre en avant la production d'un allongement vocalique et ce particulièrement dans les productions d'apprenants francophones de niveau B. Ce constat pourrait s'expliquer en menant une analyse perceptive qui confirmerait la réalisation d'une voyelle française et donc d'une trop grande tension. On pourrait étendre la comparaison aux durées syllabiques en prenant en compte la structure syllabique, les effets du contexte consonantique (Lindblom, 1963), les différences entre syllabes accentuées et inaccentuées (Gut, 2003), ainsi que le contexte prosodique. L'enjeu pédagogique est important car les résultats fournissent une aide précieuse quant à l'établissement de caractéristiques typologiques de strates d'interlangue. Une réplique de l'étude sur d'autres corpus tels que le corpus d'apprenants ANGLISH (Tortel, 2008) permettrait de confirmer ou infirmer cette possible « systématisation » de déviations caractéristiques d'un niveau de l'apprenant.
Notre recherche contribue à mieux comprendre la réalisation des voyelles inaccentuées chez les natifs anglophones et chez les apprenants, à poser une norme de référence pour la prononciation et à en établir les enjeux pédagogiques pour l'apprentissage et l'enseignement de l'anglais.

# Références

BATES S. (1995). *Towards a definition of schwa: an acoustic investigation of vowel reduction in English*. Ph.D. dissertation, University of Edinburgh.

BIGI B. (2012). SPPAS: A Tool for the Phonetic Segmentation of Speech. In *Proceedings of the Language Resource and Evaluation Conference*. Istanbul: Turkey: 1748-1755.

BOERSMA P., WEENINK D. (2001). PRAAT, a system for doing phonetics by computer. *Glot International* 5/9-10: 341-345. http://www.praat.org, Version 5.4.04

BOLINGER D. (1981). *Two Kinds of Vowels, Two Kinds of Rhythm*, Bloomington, IN, USA : IULC Publications, 68 p.

BROWMAN C., GOLDSTEIN L. (1992). "Targetless" schwa: an articulatory analysis. In LADD D.R. & DOCHERTY G. (Eds.). *Papers in Laboratory Phonology II*. Cambridge: CUP, 26-67.

CHAN D., FOURCIN A., GIBBON D., GRANDSTRÖM B., HUCKVALE M, KOKKINAKIS G., KVALE K., LAMEL L., LINDBERG B., MORENO A., MOUROPOULOS J., SENIA F., TRANSCOSO I., VELT C., ZEILIGER J. (1995). "EUROM – A Spoken Language Ressource for the EU." *In Proceedings of Eurospeech'95,* Madrid.

DETERDING D. (1997). The formants of monophthong vowels in Standard Southern British English pronunciation. *Journal of the International Phonetic Association* 27, 47-55.

ERICSDOTTER C., ERICSSON A.M. (2001). Gender differences in vowel duration in read Swedish: Preliminary results. In *Proc. Fonetik 2001, XIVth Swedish Phonetics Conference. Working Papers of the Department of Linguistics* Volume 49. Lund University, 34–37.

FLEMMING E. (2009). The phonetics of schwa vowels, in MINOVA D. (Ed.). *Phonological Weakness in English*. Houndsville : Palgrave Macmillan, 78-95.

GENDROT C., ADDA-DECKER M. (2007). Impact of duration and vowel inventory size on Formant values of aeral vowels: an automated formant analysis from eight languages. In: International Congress of Phonetics Sciences, pp.1417-1420, Germany.

GUT U. (2003) Non-native Speech rhythm in German. *Proceedings of 15th International Congress of Phonetic Sciences*, Barcelona. 2437-2440.

GUT U. (2009). *Non-native Speech. A corpus-based Analysis of Phonological and phonetic Properties of L2 English and German*, Frankfurt: Peter Lang.

HERMENT S. (2010). The pedagogical implications of variability in transcription, the case of [i] and [u]. In HENDERSON A. (Ed.). *English Pronunciation: Issues and Practices (EPIP): Proceedings of the First International Conference*. Université de Savoie, Chambéry, France, 177-188.

HERMENT S., LOUKINA A., TORTEL A. (2012). *AIXOX*. SLDR, http://sldr.org/sldr000784/fr

HERMENT S., TORTEL A., BIGI B., HIRST D., LOUKINA A. (2014). AixOx, a multi-layered learners' corpus: automatic annotation. In DÍAZ PÉREZ J., DÍAZ NEGRILLO A. (Eds.). *Specialisation and variation in language corpora*. Bern: Peter Lang, 41-76.

HERMENT S., TURCSAN G. (Soumis). Revisiting contrasts for high vowels in English: Allophony, Neutralisation and Phonological corpora.

HERRY-BENIT N., NIKOLOV R., TORTEL A. (2009). Positional determination of the quality of schwa in English. *Scientific Works – Philology*, vol. 47, Book 1 Part B. Plovdiv University, Bulgaria, 247-257.

HESELWOOD B. (2007). Schwa and the Phonotactics of RP English. *Transactions of the Philological Society,* vol. 105, no. 2.

HILLENBRAND J., GETTY L.A, CLARK M.J., WHEELER K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America* 97, 3099–3111.

HIRST *et al*. (2013). Building OMProDat: an open multilingual prosodic database. *Proceedings of TRASP*, September 2013, Aix-en-Provence, France, 11-14. http://sldr.org/sldr000725

HORGUES C. (2010). *Prosodie de l'accent français en anglais et perception par des auditeurs anglophones natifs*. Thèse de doctorat, Université Paris Diderot-Paris 7.

JONES D. (1967). *Everyman's English Pronouncing Dictionary*, 13th ed. London: J.M. Dent & Sons.

JONES D. (2011). *English Pronouncing Dictionary*, 18th edition (revised by ROACH P., SETTER J., ESLING J.). Cambridge: CUP.

LINDBLOM B. (1963). On vowel reduction, *Report #29, The Royal Ins. of Tech., Speech Transmission Laboratory,* Stockholm, Sweden.

LOBANOV B.M. (1971). Classification of Russian vowels spoken by different listeners. *JASA* 49, 606-08.

PETERSON G., BARLEY H. (1952). Control Methods Used in a Study of the Vowels, *Journal Acoustical Society of America*, Vol. 24: 175-184.

SIMPSON A.P. (1998). Phonetische Datenbanken des Deutschen in der empirischen Sprachforschung und der phonologischen Theoriebildung. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverar-beitung der Universität Kiel* (AIPUK) 33.

THOMAS E.R., KENDALL T. (2007). *NORM: The vowel normalization and plotting suite.* [Online Resource: http://ncslaap.lib.ncsu.edu/tools/norm/ ]

TORTEL A. (2008). ANGLISH : base de données comparatives l1 & l2 de l'anglais lu, répété et parlé. *Travaux Interdisciplinaires du Laboratoire Parole et Langage* (TIPA), 27, 111-122.

TORTEL A. (2009). *Evaluation qualitative de la prosodie d'apprenants français : apport de paramétrisations prosodiques*. Thèse de doctorat, Aix-Marseille Université.

WELLS J.C. (2008). *Longman Pronunciation Dictionary*, 3rd edition. London: Longman.

WENK B.J., WIOLAND F. (1982). Is French really syllable-timed? *Journal of Phonetics*, 10, 193-216.

# Organisation temporelle de la parole dans la dystonie généralisée primaire

Marie-Charlotte Cuartero[1], Roxane Bertrand[1], Marie Vidailhet,[2] David Grabli[2], Serge Pinto[1]
(1) Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
(2) Institut du Cerveau et de la Moelle, APHP, Paris
Marie-charlotte.cuartero@lpl-aix.fr

## RESUME

La dystonie généralisée primaire est un trouble neurologique se caractérisant par des mouvements involontaires et une **dysarthrie** (trouble de la production de la parole) de type **hyperkinétique**. La classification perceptive des dysarthries de Darley et *al.*, (1969 a, b) montre que dans la DGP, l'altération de l'organisation temporelle de la parole se manifeste par un débit lent, des silences inappropriés et un allongement des pauses. L'objectif de notre étude est de caractériser acoustiquement la parole dans la DGP en vue de confirmer/infirmer ces résultats perceptifs. Globalement, nos résultats ne montrent aucune différence significative entre sujets contrôle et patients atteints de DGP. L'analyse par sujet montre que quatre patients parmi les 11 étudiés présentent cependant une dysrythmie avec au moins une variable temporelle altérée. Alors que la dysarthrie est considérée comme un défaut d'exécution de la parole, il serait intéressant d'explorer l'implication de dysfonctionnements non-moteurs dans l'organisation temporelle de la parole.

## ABSTRACT

**Temporal organization of speech in primary generalized dystonia**
Primary generalized dystonia (PGD) is a neurologic disorder characterized by involuntary movements and associated with a hyperkinetic dysarthria (motor speech disorder). The seminal classification of dysarthrias made by Darley et *al.*, (1969 a, b) denotes that in PGD, the deterioration of speech temporal organization is associated with a slow rate, inappropriate silences and a pause widening. The aim of our study is to describe acoustically speech in PGD in order to confirm (or not) the previous perceptual findings. Globally, our results do not show any significant difference between healthy controls and patients with PGD. The analysis-by-subject demonstrates that 4 patients among the 11 studied however present with a dysrhythmia, with at least one temporal variable altered. While dysarthria is considered a motor speech disorder, it should be interesting to explore the contribution of non-motor dysfunctions in speech temporal organization.

MOTS-CLES : prosodie, organisation temporelle, dystonie généralisée primaire, dysarthrie.
KEYWORDS : prosody, temporal organization, primary generalized dystonia, dyarthria.


10.21437/JEP.2018-68

# 1 Introduction

## 1.1 La dystonie généralisée primaire

La dystonie est un trouble neurologique défini par « des contractions musculaires soutenues ou intermittentes, provocant des mouvements et/ou des postures anormales » (Albanese *et al.*, 2013). Se caractérisant par des mouvements involontaires, elle est classée dans les **troubles hyperkinétiques** du mouvement. La cause de ce désordre est due à un dysfonctionnement des noyaux gris centraux. La dystonie généralisée peut être considérée comme un syndrome, et dans ce cas elle est considérée comme « primaire », à la différence des dystonies « secondaires » qui se manifestent généralement plus tardivement, à la suite d'un évènement spécifique. L'origine de la dystonie généralisée primaire (DGP) est inconnue ou génétique, la forme DYT1 est la plus sévère et la plus connue (Albanese *et al.*, 2013). Elle débute au niveau d'un membre puis tend à se généraliser au cours du temps (Vercueil, 2007). Elle se développe dès l'enfance et survient avant 26 ans. La prévalence de la dystonie à début précoce est de 2 à 50 cas par million (Defazio et *al.*, 2010).

## 1.2 La dysarthrie hyperkinétique dans la dystonie généralisée primaire

La **dysarthrie hyperkinétique** est un symptôme présent dans la DGP. C'est un trouble de l'exécution motrice de la parole pouvant affecter la phonation, l'articulation et la prosodie (Darley et *al.*, 1969 a, b). Les quatre grandes *caractéristiques* de la parole dans la DGP sont : **l'imprécision articulatoire**, **la sténose phonatoire, l'insuffisance et/ou l'excès prosodiques** (Darley et *al.*, 1969 a, b ; cf. table 1). La parole dans la DGP se manifeste ainsi par une articulation et une intelligibilité dégradée, une voix rauque, forcée avec des variations excessives d'intensité et des arrêts vocaux. Ces altérations entraînent la production de phrases courtes avec des inspirations très perceptibles. Les distorsions phonatoires et articulatoires génèrent des troubles prosodiques rendant notamment le débit de parole lent (Tripoliti, 2007). Dans la définition de la dysarthrie, ces dysfonctionnements sont regroupés sous le terme de « dysprosodie » (Duez, 2007). Ils peuvent se caractériser par des faits de dysmélodie (distorsions de l'intonation) et/ou de dysrythmie (altération de l'organisation temporelle) (Duez, 2007). Cette dernière renvoie à une diminution des marques de proéminences, un débit lent, des silences inappropriés, des arrêts vocaux, un allongement des pauses et des phonèmes qui participent à la dysprosodie dans la DGP (Darley et *al.*, 1969 a, b).

| | |
|---|---|
| Sténose phonatoire | Voix rauque<br>Voix forcée<br>Arrêts vocaux<br>Variation excessive d'intensité |
| Imprécision articulatoire | Articulation dégradée<br>Distorsion des voyelles<br>Imprécision des consonnes |
| Excès prosodique | Débit lent<br>Silences inappropriés<br>Allongements des pauses<br>Allongement des phonèmes |
| Insuffisance prosodique | Monotonie<br>Mono-intensité<br>Phrases courtes<br>Diminution de l'accentuation |

TABLE 1. Caractéristiques de la dysarthrie hyperkinétique dans la DGP (selon Darley et al., 1969 a, b).

### 1.3 Problématique et hypothèses

La description de référence réalisée par Darley et *al.*, (1969 a, b) repose sur des analyses perceptives. Comme Darley et *al.*, d'autres travaux ont observé que les paramètres identifiés comme déviants perceptivement étaient difficiles à distinguer entre la dysarthrie dans la DGP et d'autres types de dysarthries (Zyski et Weisiger, 1987). De nouvelles analyses acoustiques sont donc cruciales pour identifier de manière objective les caractéristiques de cette dysarthrie (Kent et *al.*, 1999). Dans ce sens, des études antérieures ont mis en évidence que le débit respiratoire rapide, les moments d'apnée associés à des dysrythmies et la baisse du volume respiratoire définiraient plus spécifiquement la dysarthrie hyperkinétique (LaBlance et Rutherford, 1991). Notre objectif ici est d'approfondir nos connaissances sur la dysarthrie dans la DGP en caractérisant mieux sur le plan acoustique, les paramètres altérés.
Largement inspiré des travaux de Duez (2007), nous nous attendons à observer des paramètres définissant une dysrythmie et une désorganisation temporelle dans les productions de parole des patients atteints de DGP. En revanche, compte tenu de l'hétérogénéité des patients, ces dysfonctionnements devraient plutôt apparaître chez les patients dont le degré de sévérité de la dysarthrie est important (Tripoliti, 2007).

## 2 Matériels et méthode

### 2.1 Participants

Onze patients présentant une DGP ont été recrutés dans les centres hospitaliers universitaires de Grenoble, Paris, Nantes et Bordeaux (France). La qualité de leur voix a été évaluée par une orthophoniste à l'aide d'une évaluation clinique du degré de sévérité de la voix (G=grade général de dysphonie, R=raucité, B=souffle, A=asthénie, S=serrage laryngé [GRBAS] ; Hirano,1981). L'ensemble des données démographiques et cliniques sont disponibles dans le Tableau 1. Les patients ont été sous traitements médicamenteux optimal pour la prise en charge de leur troubles dystoniques. Aucun traitement spécifique pour la parole n'était utilisé. Tous les patients remplissaient les critères d'inclusion pour une neurochirurgie de type stimulation cérébrale profonde. Par ailleurs, onze sujets contrôles sains, appariés en âge et en genre, ne présentant aucun trouble de la voix et aucun antécédent neurologique, ont participé à l'étude (8 femmes, âge moyen = 42,6, SD âge = 13,2 ; 3 hommes, âge moyen = 31,7, SD âge = 12,6). Ces derniers ont été enregistrés au Laboratoire Parole et Langage (UMR 7309, LPL, Aix-en-Provence). Au préalable l'ensemble des participants a signé un consentement de participation, dans le respect des lois éthiques en vigueur, et en accord avec la déclaration d'Helsinki (OMS, 2004).

| **Patients** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Genre** | F | F | H | F | F | H | H | F | F | F | F |
| **Age (ans)** | 38 | 24 | 30 | 40 | 35 | 20 | 44 | 35 | 50 | 56 | 66 |
| **Age au moment du diagnostic (ans)** | 7 | 6 | 10 | 20 | 10 | -- | 13 | 21 | 47 | 54 | 59 |
| **Score G de l'échelle GRBAS** | 2 | 1 | 3 | 2 | 2 | 3 | 1 | 0 | 1 | 1 | 1 |
| **Score global BFMDRS** | 73.5 | 30.5 | 29 | 23 | 64.5 | 82 | 22 | 14 | -- | 45 | -- |

TABLE 2 : Tableau clinique des patients avec Dystonie Généralisée Primaire. F = femme, H = homme ; GRBAS : 0 = voix normale, 1 = dysphonie légère, 2 = dysphonie moyenne, 3 = dysphonie sévère. Burke-Fahn-Marsden Dystonia Rating Scale (BFMDRS) : échelle d'évaluation des mouvements pour les Dystonies (minimum = 0, maximum = 120). *En raison d'une parole inintelligible, le patient 6 a été exclu des analyses (colonne grisée).*

## 2.2 Corpus

Le corpus exploité consiste en une tâche de lecture de texte (« La chèvre de Monsieur Seguin », Daudet, 1869). Les sujets ont reçu pour consigne de le lire avec un débit de parole le plus naturel possible. Les enregistrements ont été acquis à l'aide d'un microphone (modèle 444 PP, AKG) relié à un enregistreur digital micro Track 24/96, M-Audio.

## 2.3 Pré-traitement des enregistrements de la parole

### 2.3.1 Etape préalable de segmentation

Le plugin EasyAlign a été utilisé pour effectuer la segmentation en phonèmes, syllabes, mots et phrases à partir des fichiers audios et de la transcription orthographique. Les frontières temporelles et les pauses silencieuses (notées #) ont été corrigées manuellement. Le seuil minimum des pauses a été estimé à 150 millisecondes (Kent et *al.*, 1999).

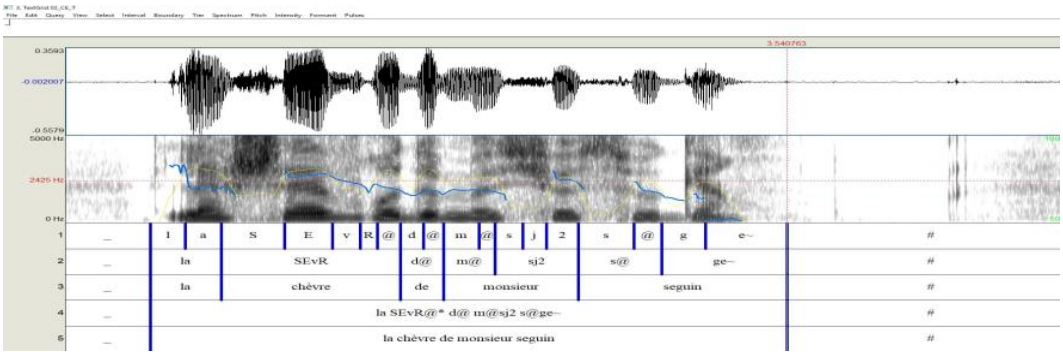



Figure 1. Segmentation en phonèmes (tier 1), syllabes (tier 2), mots (tier 3) et phrases (tier 4 et 5) d'après le plugin EasyAlign.

### 2.3.2 Variables temporelles

Les variables temporelles considérées pour l'analyse de l'organisation temporelle de la parole sont :

- Temps total de parole (en secondes) ;
- Nombre de syllabes ;
- Vitesse de parole (nombre de syllabes / temps total de parole, en secondes) ;
- Temps total d'élocution (durée totale des syllabes, en secondes) ;
- Vitesse d'élocution (nombre de syllabes / temps total d'élocution en secondes) ;
- Nombre de pauses ;
- Durée totale des pauses.

Nous avons complété l'analyse quantitative de ces variables par une analyse qualitative des variables suivantes, pour les sujets concernés :

- Localisation des pauses ;
- Disfluences : interruption du flux de parole par des répétitions et/ou des reformulations par exemple ;
- Omissions de mots.

Les pauses sont dites *syntaxiques*, lorsqu'elles marquent la frontière des paragraphes, des phrases, des propositions et des syntagmes. Elles peuvent être *non syntaxiques* lorsqu'elles sont localisées à

l'intérieur des mots, des syllabes et des syntagmes (Duez, 2007). Dans la parole pathologique, souvent marquée par une désorganisation temporelle, les pauses non syntaxiques sont plus fréquentes.
Les patients 09 et 11 n'ont lu qu'une partie du texte. Ainsi, les variables suivantes : nombre de syllabes, nombre de pauses, temps total de parole et d'élocution n'ont pu être intégrées à l'analyse pour ces patients.

## 2.4 Analyses statistiques

Les analyses statistiques ont été réalisées à l'aide du logiciel Rstudio (Version1.1.383 – © 2009-2017). Des comparaisons de moyennes ont été menées à l'aide de test t pour chaque variable temporelle. Dans un second temps, des scores standardisés (scores Z) ont été calculés pour permettre d'identifier la distance entre chacun de nos patients et l'ensemble des participants contrôles. Ce score est à interpréter comme une distance en nombre d'écart-types. Nous avons considéré qu'une distance supérieure ou égale à 2 écart-types correspondait à un résultat déviant.

## 2.5 Résultats

Aucune différence significative globale n'a été trouvée entre les patients et les contrôles pour chacune de nos variables ($p>0.05$ ; cf. tableau 2).
Sur la base de nos Scores Z, les résultats des patients 01, 02, 03 et 05 ont en revanche montré un écart aux participants contrôles supérieurs à deux écarts-types pour au moins une variable.

| DGP | Nombre de syllabes | Nombre de pauses | Durée pauses | Temps d'élocution | Temps de parole | Débit de parole | Débit d'élocution |
|---|---|---|---|---|---|---|---|
| **01** | 233 | 42 | 20,75 | 60,00 | 80,75 | 2,89 | 3,88 |
| **02** | 250 | 25 | 9,18 | 45,67 | 54,85 | 4,56 | 5,47 |
| **03** | 226 | 40 | 18,89 | 59,22 | 78,11 | 2,89 | 3,82 |
| **04** | 241 | 29 | 14,69 | 47,52 | 62,20 | 3,87 | 5,07 |
| **05** | 255 | 53 | 34,11 | 59,25 | 93,36 | 2,73 | 4,30 |
| **07** | 236 | 24 | 11,33 | 45,97 | 57,30 | 4,12 | 5,13 |
| **08** | 239 | 21 | 7,74 | 41,60 | 49,34 | 4,84 | 5,75 |
| **09** | - | - | - | - | - | 4,05 | 5,19 |
| **10** | - | - | - | - | - | 3,23 | 4,53 |
| **11** | 234 | 34 | 12,79 | 52,45 | 65,24 | 3,59 | 4,46 |
| **Contrôles** | 236 | 25 | 12,32 | 48,72 | 61,04 | 3,94 | 4,89 |

TABLE 2 : Données des variables temporelles de la parole pour chaque patient et moyenne des participants contrôles sains. Les lignes grisées correspondent aux patients présentant un écart supérieur à 2 écarts-types pour au moins une variable.

Pour les 4 patients présentant un profil de parole atypique, le nombre de pauses est plus important chez les patients ***01*** (zscore = 2,31), ***03***(zscore = 2,04) et ***05*** (zscore : 3,82). Ce dernier a aussi une durée de pauses plus longues (zscore : 5,04). Parmi leurs pauses, quasiment la moitié sont *non syntaxiques*, et peuvent être associées à des disfluences :

- Patient ***01*** : 42 pauses dont 19 *non syntaxique*.
- Patient ***03*** : 40 pauses dont 17 *non syntaxique*.
- Patient ***05*** : 53 pauses dont 33 *non syntaxique* (cf. exemple 1).

Le patient ***03*** a une baisse du nombre de syllabes (zscore = - 4,49), associée à une omission de 10 mots. A l'inverse, les patients ***02*** et ***05*** ont un nombre de syllabes supérieur (zscore = 5,39 ; 3,82) à

celui des locuteurs contrôles, correspondant respectivement à 14 et 19 syllabes répétées, associées à des disfluences (pour ces deux patients, n=7).

Le débit de parole et d'élocution est plus lent chez les patients ***01*** (zscore = -2,02 ; -2,63), ***03*** (zscore = -2 ; -2,80) et ***05*** (zscore : -2,31).
Le temps de parole et d'élocution est plus long chez les patients ***01*** (zscore = 2,57 ; 2,83), ***03*** (zscore = 2,22 ; 2,64) et ***05*** (zscore : 4,21 ; 2,65).

Exemple 1. Transcription d'un extrait du patient ***05*** :

Les chè- # (pause : 0.478 sec, *non syntaxique*) les chèvres s'ennuient chez moi # (pause : 0.409 sec) je n'en garderai # (pause : 0.698 sec, *non syntaxique*) pas # (pause : 0.490 sec, *non syntaxique*) une # (pause : 0.492 sec, *non syntaxique*) pas une # (pause : 1.308 sec) cependant # (pause : 0.390 sec) il ne # (pause : 0.350 sec, *non syntaxique*) il ne se découragea # (pause : 0.376 sec, *non syntaxique*) pas.

# 3 Discussion

L'objectif de cette étude est d'apporter une description acoustique de l'organisation temporelle de la parole dans la DGP. Si les résultats sur l'ensemble des sujets montrent qu'il n'existe aucune différence significative entre les participants contrôles sains et les patients présentant une DGP, l'examen individuel des patients révèle des profils dysarthriques spécifiques. Ainsi, quatre patients présentent une altération importante d'au moins un paramètre rythmique de la parole. Le profil atypique de ces patients est en lien avec l'évaluation clinique du degré de sévérité de la voix (TABLE 2). L'analyse perceptive de Darley et *al.*, (1969, a, b) a caractérisé la dysprosodie dans la DGP comme un excès et une insuffisance prosodique. Nos principaux résultats, relatifs à l'organisation temporelle montrent que le nombre, la durée des syllabes et des pauses ainsi que la localisation de celles-ci sont la conséquence d'une bradylalie (extrême lenteur) (Duez, 2007). Selon Duez (2007), ces modifications de la parole seraient caractérisées par une exagération des patterns rythmiques.

## 3.1 Exagération de patterns rythmiques

Les patients ont un débit lent, un temps de parole et d'élocution plus long. Cette manifestation bradylalique évoque une insuffisance prosodique. Ces faits de dysrythmie montrent que l'altération de la parole dans la DGP se manifeste par une hyperrythmie. Par ailleurs, LaBlance et Rutheford (1991) ont mis en évidence un débit respiratoire rapide, des moments d'apnée associés à des dysrythmies et une baisse du volume respiratoire dans la DGP. Nos résultats sont en accord avec ces manifestations physiologiques, dans le sens où elles peuvent contribuer, du moins partiellement, aux altérations de débit de parole et d'élocution.

## 3.2 Désorganisation temporelle de la parole

Par ailleurs, pour les patients les plus atteints, la proportion de pauses non syntaxiques a tendance à augmenter. Ces pauses peuvent être associées à des disfluences telles que des répétitions et/ou des reformulation de mots qui contribuent à augmenter le nombre de syllabes tandis que les omissions de mots, à l'inverse, contribuent à le baisser. Ces caractéristiques peuvent rendre compte d'une modification de la structure superficielle et d'une désorganisation temporelle de la parole (Duez, 2007). Ainsi, le profil des pauses dans la parole pathologique pourrait être la marque de troubles moteurs comme cognitifs (Duez, 2007). Les patterns de pauses observées chez nos patients pourraient

être le reflet d'altérations non exclusivement restreintes à des répercussions motrices de la DGP. Leur position, considérée comme non syntaxique, pourrait être un indicateur de déficits langagiers.

### 3.3 Modélisation de la production de la parole

La définition de la dysarthrie correspond à un déficit de l'exécution motrice de la parole, associé à une lésion du système nerveux central et/ou périphérique (Darley et *al.*, 1969, a, b). Souvent la dysarthrie n'est étudiée que sous les aspects moteurs. Pourtant, notre étude montre une désorganisation temporelle de la parole dans la DGP. Ces faits de dysrythmie ne semblent pas être exclusivement liés à des altérations motrices. En effet, d'autres études ont progressivement considéré l'implication des aspects non-moteurs dans la production de la parole. Le modèle GODIVA (Gradient Order Directions Into Volicities of Articulators, Guenther, 1994 ; Guenther & Hickok, 2015) intègre une boucle d'exécution et de planification motrice. Une des originalités de ce modèle est la prédiction de la nature et de la position des disfluences, ce qui est un aspect important pour améliorer le traitement automatique de la parole dysarthrique (Laaridh et al., 2016). Le modèle ACT (Vocal Tract ACTion), quant à lui, considère les noyaux gris centraux comme impliqués dans le contrôle et la correction d'actions motrices (Kröger et al., 2009). En ce qui concerne les modèles neurolinguistiques du rythme, il est aussi important de considérer le modèle SEP (Sound Enveloppe Processing et Synchronization and Entrainment to a pulse ; Fujii & Wan, 2014), qui attribue aux noyaux gris centraux une place importante dans la production du rythme de la parole.

## 4 Conclusion et perspectives

Afin d'enrichir cette description de la dysarthrie hyperkinétique dans la DGP, il est nécessaire de proposer une analyse approfondie de la prosodie. Dans ce cadre, il faut tenir compte de l'hétérogénéité inter-individuelle et d'avoir recours à une analyse acoustique individuelle de la parole. En effet, la variabilité de l'atteinte de la parole peut s'expliquer par la sévérité du trouble, l'évolution de la maladie et la localisation du dysfonctionnement neurologique (Tripoliti, 2007). Par ailleurs, alors que la dysarthrie est considérée comme un défaut d'exécution motrice de la parole, il serait intéressant d'investiguer l'implication d'autres fonctions comme la planification de la parole. Les répercussions de la dysarthrie devraient-elles être restreintes à des altérations motrices ou bien impliquer également des altérations non-motrices ? Notre étude cherche à apporter des éléments de réponse à cette question fondamentale et notre contribution acoustique bien que préliminaire est essentielle dans la mesure où il existe très peu d'études sur la parole dans la DGP. Pour aller plus loin, nous avons pour objectif d'étudier l'organisation des proéminences, la localisation des pauses et des disfluences au sein des syntagmes accentuels. Ils sont autant d'indicateurs susceptibles d'être altérés par la DGP.

# Références


Albanese, A., Bhatia, K., Bressman, S. B., Delong, M. R., Fahn, S., Fung, V. S. C., … Teller, J. K. (2013). Phenomenology and classification of dystonia: A consensus update. *Movement Disorders*, *28*(7), 863–873. https://doi.org/10.1002/mds.25475

Burke, R., Fahn, S., Marsden, CD., Bressman, SB., Moskowitz, C., Friedman, J. (1985). Validity and reliability of a rating scale for the primary torsion dystonias. *Neurology*, 35(1), 73 - 77.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969, a, b). Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research*, *12*(2), 246–269. http://doi.org/10.1044/jshr.1202.246

Defazio, G. (2010). The epidemiology of primary dystonia: Current evidence and perspectives. *European Journal of Neurology*, *17*(SUPPL. 1), 9–14. https://doi.org/10.1111/j.1468-1331.2010.03053.x

Duez, D (2007). Proposition pour une typologie et une évaluation acoustique des faits de dysprosodie. Dans Auzou P., Rolland-Monnoury V., Pinto, S. & Ozsancak C., (dir.), *Les dysarthries* (1ère ed., vol. 1, p. 270-279). Marseille : Solal.

Duez, D (2007). Prosodie et rythme. Dans P. Auzou, V. Rolland-Monnoury, C. Ozsancak, (dir.), *Les dysarthries* (1ère éd., vol. 1, p. 181-188). Marseille: Solal.

Fujii, S., & Wan, C. Y. (2014). The Role of Rhythm in Speech and Language Rehabilitation: The SEP Hypothesis. *Frontiers in Human Neuroscience*, *8*(October), 1–15. http://doi.org/10.3389/fnhum.2014.00777

Galaz, Z., Mekyska, J., Mzourek, Z., Smekal, Z., Rektorova, I., Eliasova, I., … Berankova, D. (2015). Prosodic analysis of neutral, stress-modified and rhymed speech in patients with Parkinson's disease. *Computer Methods and Programs in Biomedicine*, *127*, 301–317. http://doi.org/10.1016/j.cmpb.2015.12.011

Guenther, F. H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics*, *72*(1), 43–53. http://doi.org/10.1007/BF00206237

Guenther, Frank H. (2015). Role of the auditory system in speech production. *Handbook of clinical neurology*, *129*, 161–175. https://doi.org/10.1016/B978-0-444-62630-1.00009-3

Hendrix, C. M., & Vitek, J. L. (2012). Toward a network model of dystonia. *Annals of the New York Academy of Sciences*, *1265*(1), 46–55. https://doi.org/10.1111/j.1749-6632.2012.06692.x

Hirano, M. (1981). Psycho-acoustic evaluation of voice: GRBAS scale for evaluating the hoarse voice. *Clinical Evaluation of Voice*, Springer Verlag, Wien.

Kent, R. D., Weismer, G., Kent, J. F., Vorperian, H. K., & Duffy, J. R. (1999). Acoustic studies of dysarthric speech: Methods, progress, and potential. *Journal of Communication Disorders*, *32*(3), 141–186. https://doi.org/10.1016/S0021-9924(99)00004-0

Kröger, B. J., Kannampuzha, J., & Neuschaefer-Rube, C. (2009). Towards a neurocomputational model of speech production and perception. *Speech Communication*, *51*(9), 793–809. http://doi.org/10.1016/j.specom.2008.08.002

LaBlance, G. R., & Rutherford, D. R. (1991). Respiratory dynamics and speech intelligibility in speakers with generalized dystonia. *Journal of Communication Disorders*, *24*(2), 141–156. https://doi.org/10.1016/0021-9924(91)90018-E

Laaridh, I., Fredouille, C., & Meunier, C. (2016). Détection automatique d'anomalies sur deux styles de parole dysarthrique : parole lue vs spontanée. *J.E.P. 2016*, *1*(August), 229–237.

Lehericy, S., Tijssen, M. A. J., Vidailhet, M., Kaji, R., & Meunier, S. (2013). The anatomical basis of dystonia: Current view using neuroimaging. *Movement Disorders*, *28*(7), 944–957. https://doi.org/10.1002/mds.25527

Ricciardi, L., Ebreo, M., Graziosi, A., Barbuto, M., Sorbera, C., Morgante, L., & Morgante, F. (2016). Speech and gait in Parkinson's disease: When rhythm matters. *Parkinsonism and Related Disorders*, *32*, 42–47. http://doi.org/10.1016/j.parkreldis.2016.08.013

Skodda, S., Flasskamp, A., & Schlegel, U. (2010). Instability of syllable repetition as a model for impaired motor processing: Is Parkinson's disease a "rhythm disorder"? *Journal of Neural Transmission*, *117*(5), 605–612. http://doi.org/10.1007/s00702-010-0390-y

Tripoliti, E. (2007). Parole et dystonies. Dans P. Auzou, V. Rolland-Monnoury, C. Ozsancak, (dir.), *Les dysarthries* (1ère éd., vol. 1, p. 415-421). Marseille : Solal.

Vercueil, L. (2007). Les dystonies. Dans P. Auzou, V. Rolland-Monnoury, C. Ozsancak, (dir.), *Les dysarthries* (1ère éd., vol. 1, p. 407-414). Marseille: Solal.

Vitek, J. L. (2002). Pathophysiology of dystonia: A neuronal model. *Movement Disorders*, *17*(S3), S49–S62. https://doi.org/10.1002/mds.10142

World Medical Association General Assembly, Declaration of Helsinki, Amendment, (2004).

# Perception des voyelles nasales du français par des apprenants hispanophones

David Alejandro Bustamante  Pierre Hallé  Claire Pillot-Loiseau
Laboratoire de Phonétique et Phonologie (LPP) UMR7018, CNRS-Paris3 / Sorbonne Nouvelle
19 rue des Bernardins, 75005 Paris, France
david.alejandro.bustamante@gmail.com; pierre.halle@univ-paris3.fr
claire.pillot@sorbonne-nouvelle.fr

RESUME

Notre objectif est d'examiner la perception des voyelles nasales françaises par des apprenants hispanophones d'Espagne et de Colombie de divers niveaux grâce à une série de tests de catégorisation et de discrimination. Les résultats de catégorisation des voyelles nasales du français (choix forcé parmi /ɛ̃, ɑ̃, ɔ̃/) dans des logatomes dissyllabiques dans les positions initiale, médiale et finale, montrent que les apprenants hispanophones ont 1) plus de difficulté pour /ɛ̃/ et /ɑ̃/ pour lesquelles ils donnent des réponses très variables ; 2) moins de difficulté avec /ɔ̃/, surtout en position finale (finale>médiale>initiale). Un test de catégorisation avec choix forcé entre les cinq catégories vocaliques de l'espagnol (/a, i, u, e, o/) permet de mieux comprendre comment les sujets hispanophones naïfs assimilent les voyelles nasales françaises. Nous proposons une modélisation quantitative, basée sur PAM (Best, 1995), prédisant les performances de discrimination à partir des assimilations et des notes de confiance subjective.

ABSTRACT

**Perception of Spanish-speaking learners of French nasal vowels**

The present study is intended to examine how Spanish-speaking learners of French from Spain and Colombia perceive the French nasal vowels according to level of fluency, using a series of categorization and discrimination. The results of the forced-choice within /ɛ̃, ɑ̃, ɔ̃/ categorization tests on French nasal vowels appearing in two-syllable non-words in initial, medial, or final position show that Spanish-speaking learners encounter a greater difficulty with /ɛ̃/ and /ɑ̃/, which receive quite variable responses; they encounter a lesser difficulty with /ɔ̃/, especially in word-final position (final>medial>initial). A further categorization test requiring forced-choice responses within the five Spanish vowel categories (/a, i, u, e, o/) helped to better understand how Spanish-speaking listeners naive with respect to French assimilate French nasal vowels. We propose a quantitative modeling, based on PAM (Best, 1995), whereby discrimination performance is predicted from the observed assimilation and goodness-of-fit ratings.

MOTS-CLES : perception de L2, voyelles nasales du français, apprenants hispanophones espagnols et colombiens, sujets naïfs

KEY WORDS: L2 perception, French nasal vowels, Spanish and Colombian learners, naïve subjects


10.21437/JEP.2018-69

# 1. Introduction

La prononciation d'une langue étrangère pose des difficultés d'apprentissage aux adultes : beaucoup de travaux se sont intéressés à ce sujet. Depuis la notion de « crible phonologique » (Troubetzkoy, 1938/1967), où le système phonologique de la langue maternelle agit comme un filtre influençant la perception des sons non natifs et par conséquent leur reproduction, d'autres travaux plus récents se sont attachés à rendre compte de ce phénomène concernant la perception des sons non-natifs par des auditeurs naïfs comme le *Modèle d'assimilation perceptive* (*Perceptual Assimilation Model*, PAM ; Best, 1995), ou par des apprenants d'une seconde langue (voir aussi PAM-L2 : Best & Tyler, 2007). Selon le PAM, l'assimilation des sons d'une seconde langue (L2) à des catégories de la langue native (L1) se produit non seulement dans la perception des sons de manière isolée mais aussi dans la perception des contrastes de la L2. Un contraste de la L2 pouvant être assimilé à deux catégories différentes de la L1 (*two category TC assimilation type*) est mieux acquis qu'un contraste non-natif assimilé à une même catégorie de la L1 (*single category SC assimilation type*), ou bien un segment du contraste est acceptable et l'autre est déviant (*category goodness CG assimilation type*).

Le *Modèle d'apprentissage de la parole* (*Speech Learning Model*, SLM : Flege, 1995) représente une autre référence pour la perception des sons non-natifs mais il se concentre plus sur la production que la perception. D'autres travaux sur la perception de la L2 sont le *Modèle magnétique de la langue native* (*Native Language Model*, NLM : Kuhl, 2000), qui étudie notamment la perception de la parole chez des jeunes enfants à travers la notion d'aimants perceptifs constituant des prototypes des catégories des sons du langage ; par rapport aux apprenants d'une L2, les sons non-natifs sont attirés par des aimants de la L1. Enfin, le *Modèle de perception linguistique d'une seconde langue* (*Second Language Linguistic Perception Model*, L2LP : Escudero, 2005), rend compte de l'influence que porte l'expérience de la L1 sur la perception et l'apprentissage d'une L2 par des sujets naïfs, des apprenants débutants à très avancés.

A notre connaissance, il existe très peu d'études sur la perception des voyelles nasales du français. Le travail d'Inceoglu (2014) aborde une perspective comparative de la perception des voyelles nasales du français après un entraînement multimodal (audio-visuel, visuel, et audio) auprès d'apprenants américains. Les effets de l'entraînement indiquent une progression de l'identification ([ɔ̃] > [ɛ̃] > [ɑ̃]) par rapport à l'expérience pré-entraînement, surtout pour [ɛ̃], dans toutes les modalités soulignant la contribution des gestes visuels. Detey et al. (2015) présente les résultats d'une étude longitudinale sur le rapport entre perception et production des voyelles nasales du français par des apprenants japonais ; les résultats d'une expérience de discrimination montrent que l'opposition /ɑ̃/-/ɛ̃/ est mieux perçue que l'opposition /ɑ̃/-/ɔ̃/ par ces apprenants, mais que le taux de discrimination par voyelle est plus bas pour /ɛ̃/ lorsqu'il est opposé à /ɑ̃/. Une autre étude sur la perception et la production des voyelles nasales auprès d'apprenants chypriotes hellénophones, montre la capacité des apprenants à bien distinguer le trait de nasalité du trait oral dans les voyelles du français, mais souligne la confusion du timbre entre /ɛ̃/ et à /ɑ̃/ tant en perception qu'en production (Kakoyianni-Doa et al. 2017). D'autres études portent sur la perception des voyelles nasales du français par des apprenants japonophones (Kamiyama, 2009), arabophones jordaniens (Nawafleh, 2013), et brésiliens (Desmeules-Trudel, 2013).

Par ailleurs, les recherches sur la perception des hispanophones concernant le système vocalique du français sont très limitées ; nous trouvons seulement l'étude de Magnen et al. (2005) sur la perception de /i/ /y/ /u/ du français. Nous constatons qu'il n'y a pas de recherches sur la perception

des voyelles nasales du français par des hispanophones. C'est pourquoi nous menons cette étude, mais aussi parce que d'autres travaux sur les voyelles nasales des hispanophones en production ont montré la difficulté qu'elles représentent pour l'apprentissage de la prononciation du français (Detey et al. 2010 ; Bustamante et al. 2014). En espagnol, le trait de nasalité existe pour les consonnes, mais il n'y a pas de voyelles nasales, seulement une nasalisation des voyelles par coarticulation.

L'objectif de cette étude est d'examiner la perception des voyelles nasales du français par deux groupes d'hispanophones : espagnols et colombiens. Le motif des deux populations d'hispanophones est d'étudier s'il y a une différence de traitement de la perception des voyelles nasales dans ces deux variétés de l'espagnol, tant par les apprenants que par les sujets naïfs du français. Nous avons effectué un premier test de catégorisation des voyelles nasales dans des non-mots à deux syllabes, afin d'observer si la position syllabique de la voyelle joue un rôle dans la perception. Sur la base des résultats de la première expérience, nous avons ensuite effectué un test de catégorisation des voyelles nasales selon les catégories vocaliques de l'espagnol avec des sujets espagnols et colombiens naïfs pour ce qui concerne le français pour observer les types d'assimilation des voyelles nasales. Finalement, basés sur les principes de PAM (Best, 1995), nous proposons une modélisation prédisant les performances de discrimination des voyelles nasales du français par les apprenants et les naïfs hispanophones.

# 2. Matériel et méthode

## 2.1 Expériences et stimuli

Nous avons mené trois types d'expériences pour étudier la perception des trois voyelles nasales françaises, /ɛ̃/, /ɑ̃/, /ɔ̃/, présentes dans la région métropolitaine de Paris, par les apprenants et les sujets naïfs hispanophones.

La première expérience a consisté en un test d'identification à choix forcé des voyelles nasales à choix forcé dans des non-mots à deux syllabes, test effectué auprès des apprenants espagnols et colombiens. L'intérêt d'utiliser des logatomes à deux syllabes est de nous permettre d'observer l'identification des voyelles nasales en fonction de trois positions syllabiques différentes, présentées séparément : position initiale absolue #_CCV, position initiale post-consonantique CC_CV, et position finale CVC_#. 108 stimuli ont été élaborés pour former les logatomes, produits par deux francophones natifs : un homme (54 stimuli), et une femme (54 stimuli) de la région Île-de-France. Le nombre des stimuli par sous-test, c'est-à-dire par position syllabique de la voyelle nasale est de 36. Le contexte phonétique des stimuli par sous-test est : [#_pru] ; [bl_tra, gl_se] ; [claf_#, vap_#].

Ultérieurement, nous avons effectué un test d'identification des catégories phonétiques natives auprès de sujets hispanophones naïfs du français, avec le but d'observer les types d'assimilation que ces sujets réalisent des voyelles nasales dans leur système vocalique natif. Les stimuli critiques sont à 18 mots monosyllabiques du français portant une voyelle nasale. Pour le contexte phonétique, nous avons : [fɛ̃, fɔ̃, sɛ̃, sɑ̃, sɔ̃, pɑ̃, pɔ̃, tɛ̃, tɑ̃, tɔ̃, vɛ̃, vɑ̃, vɔ̃, mɛ̃, mɑ̃, lɛ̃, lɑ̃, lɔ̃]. Ils ont été produits par deux locuteurs francophones natifs de la région de l'Île-de-France (une femme et un homme), donc 36 stimuli au total. Une erreur technique ne nous a permis d'équilibrer le contexte phonétique avec [fɑ̃], [pɛ̃], et [mɔ̃].

Enfin, les apprenants et les sujets naïfs ont passé un test de discrimination AXB sur les voyelles nasales du français. L'objectif de cette expérience est de prédire quelles sont les oppositions des voyelles nasales du français difficiles à discriminer par les auditeurs hispanophones. Douze mots monosyllabiques du français ont été utilisés pour former les stimuli du test de discrimination ; ils ont été produits trois fois par deux locuteurs (une femme et un homme de la région Ile-de-France) : 12 mots x 3 fois x 2 locuteurs = 72 stimuli. Le contexte phonétique des stimuli est : [fṼ, pṼ, tṼ, kṼ, vṼ, lṼ], où Ṽ est une des trois voyelles nasales.

## 2.2 Participants

30 apprenants espagnols (âge moyen 21.4 ; ET 5.2) et 26 apprenants colombiens (âge moyen 20.2 ; ET 2.3) ont participé à l'étude de perception. Nous les avons classés en fonction de leur niveau de français selon leur temps d'apprentissage et selon leur niveau dans l'établissement où ils apprennent le français : de six mois à un an d'apprentissage, les apprenants ont été classés comme *débutants* ; de un an et demi à deux ans et demi, ils sont dans le niveau *intermédiaire* ; et de trois ans et plus, ils sont dans le niveau *avancé* (Tableau 1). Les apprenants et les sujets naïfs ont participé à cette étude depuis leur ville d'origine : à Bogota, pour les colombiens, et à Séville pour les espagnols. 18 francophones natifs (16 F, 2 H) ont également participé en tant que groupe contrôle.

| | naïfs | débutants | intermédiaires | avancés |
|---|---|---|---|---|
| Espagnols | **25** (21 F, 4 H) | **12** (10 F, 2 H) | **9** (6 F, 3 H) | **9** (5 F, 4 H) |
| Colombiens | **19** (11 F, 8 H) | **10** (6 F, 4 H) | **9** (8 F, 1 H) | **7** (5 F, 2 H) |

Tableau 1. Nombre des sujets naïfs, et des apprenants espagnols et colombiens selon leurs niveaux de français. F pour femme, H pour homme.

## 2.3 Procédure

Une phase d'entraînement pour chaque type de test a été proposée aux apprenants afin qu'ils se familiarisent avec le type de test, mais aussi pour qu'ils ce concentrent sur la position syllabique des voyelles nasales dans l'ensemble des sous-tests de catégorisation des voyelles nasales à choix forcé dans des logatomes. Les participants ont passé tout d'abord un sous-test d'identification des 36 logatomes contenant une voyelle nasale en position initiale absolue (#_CCV), où ils devaient indiquer sur le clavier d'un ordinateur quelle était la voyelle nasale entendue dans cette syllabe : /ɛ̃/, /ɑ̃/, ou /ɔ̃/. Comme la plupart des apprenants n'étaient pas très familiarisés avec les symboles phonétiques, nous avons employé une transcription orthographique pour les réponses : <in>, <an>, et <on>. Le principe était le même pour les autres sous-tests d'identification où la voyelle nasale se trouvait en position initiale post-consonantique (CC_CV), et en position finale (CVC_#).

La catégorisation des voyelles nasales selon les 5 catégories vocaliques de l'espagnol a été passée par les apprenants et les sujets naïfs hispanophones. Après avoir entendu le stimulus, les auditeurs devaient signaler sur le clavier d'un ordinateur si le son vocalique entendu correspondait à l'une des voyelles de l'espagnol : /i/, /e/, /a/, /o/, /u/. Ensuite, sur une échelle de 1 à 5, ils indiquaient le degré de bonne correspondance pour la catégorie attribuée à la voyelle nasale (*5* : bonne correspondance ; *1* : mauvaise correspondance). Finalement, nous avons effectué un test de

discrimination AXB des voyelles nasales. Après avoir entendu chaque triplet, le participant devait indiquer sur le clavier d'un ordinateur si le son correspondant à X était plutôt similaire au premier son (A), ou plutôt au troisième son (B).

# 3. Résultats

## 3.1 Identification des voyelles nasales selon la position syllabique

Les résultats d'identification des voyelles nasales ont été analysés en termes de taux d'identification correcte pour chaque population (colombiens, espagnols, et français) en fonction de la position syllabique des voyelles dans les logatomes. Les résultats montrent que, toutes voyelles et positions confondues, le groupe français présente le taux d'identification correcte le plus élevé (99%), suivi du groupe espagnol (62%), et en dernier du groupe colombien (53%). La différence entre les trois groupes est significative, $F_{(2,7989)} = 698.2$, $p < 0.0001$. Il y a aussi une interaction entre les effets Groupe et Voyelle nasale, $F_{(4,7983)} = 30.44$, $p < 0.0001$, indiquant que l'identification des voyelles nasales est réalisée différemment selon les groupes. En effet, il n'y a pas de différence significative chez les français sur l'identification, $F_{(2,2049)} = 1.46$, $p = 0.231$, mais il y a des différences significatives dans les groupes espagnol ($F_{(2,3237)} = 76.14$, $p < 0.0001$), et colombien ($F_{(2,2805)} = 82.5$, $p < 0.0001$).

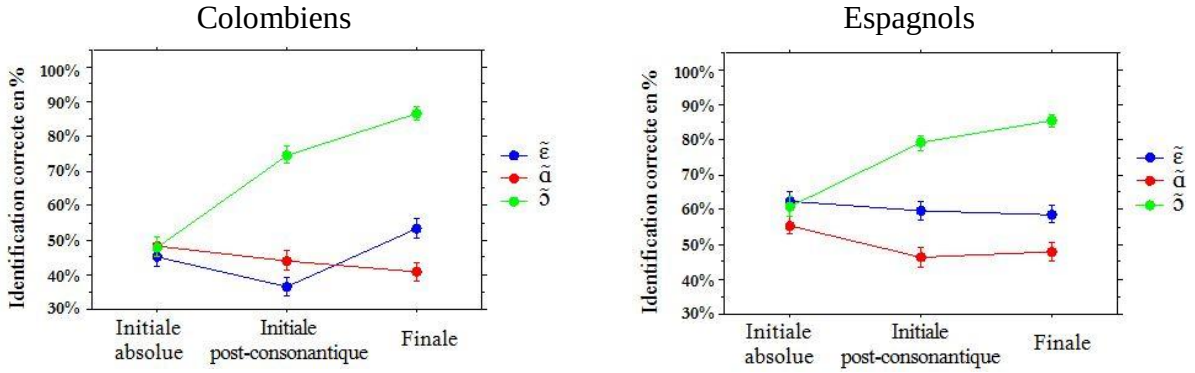



FIGURE 1. Taux moyen d'identification correcte (en pourcentages) des apprenants colombiens (à gauche) et espagnols (à droite) en fonction de la position syllabique des voyelles nasales dans le logatome : initiale absolue (#_CCV), initiale post-consonantique (CC_CV), et finale (CVC_#).

Les résultats, présentés dans la Figure 1, montrent un effet significatif concernant la position syllabique de la voyelle nasale dans le groupe colombien, $F_{(2,2805)} = 16.9$, $p < 0.0001$, ce qui n'est pas le cas dans le groupe espagnol, $F_{(2,3237)} = 2.17$ $p = 0.11$. Par contre, dans le groupe espagnol, on observe une différence significative sur l'identification entre les trois voyelles nasales (/ɔ̃>/ɛ̃/>/ɑ̃/), $F_{(2,3237)} = 76.14$, $p < 0.0001$, alors que chez les Colombiens il y a aussi une différence significative (/ɔ̃/>/ɛ̃/=/ɑ̃/), $F_{(2,2805)} = 82.5$, $p < 0.0001$, sauf qu'il y a une confusion entre /ɛ̃/ et /ɑ̃/. L'influence de la position syllabique est déterminante pour la catégorisation de /ɔ̃/, commune aux deux groupes d'apprenants : il y a une progression de la position initiale absolue à la position finale. Comme le traitement perceptif de /ɛ̃/ et /ɑ̃/ est plus difficile pour les apprenants, une variabilité d'identification est observée en fonction de la position syllabique. Nous allons observer la catégorisation que réalisent les sujets naïfs hispanophones des voyelles nasales du français selon les catégories vocaliques de l'espagnol.

### 3.2 Catégorisation des voyelles nasales selon les catégories de l'espagnol

Les résultats, présentés dans la Figure 2, montrent la catégorisation des voyelles nasales que réalisent les hispanophones selon les catégories vocaliques de leur langue maternelle. Tout d'abord, nous observons les cas d'assimilation des sujets naïfs, car l'expérience linguistique du Français par les apprenants peut influencer leur catégorisation comme nous le verrons plus bas.

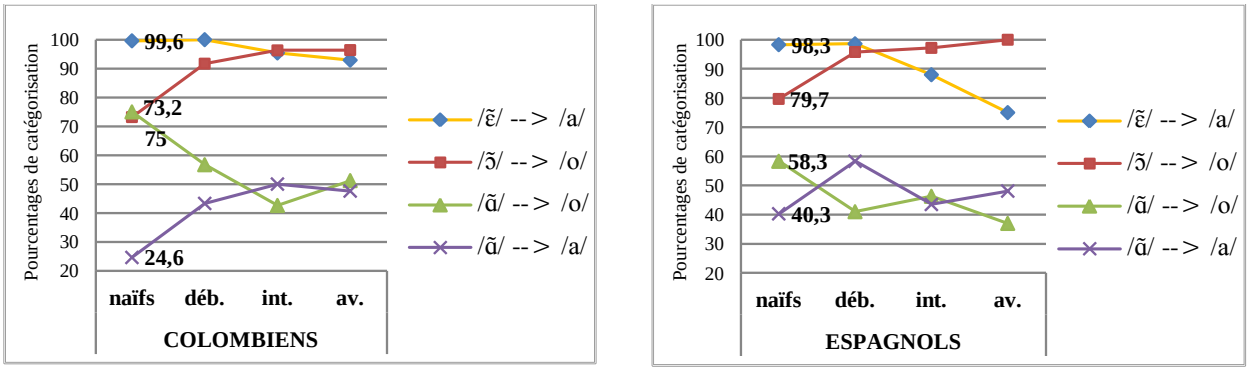



FIGURE 2. Assimilation en pourcentages des voyelles nasales du français /ɛ̃, ɑ̃, ɔ̃/, dans les catégories vocaliques de l'espagnol par les naïfs et les apprenants colombiens (gauche) et espagnols (droite).

Les sujets naïfs colombiens et espagnols assimilent /ɛ̃/ à la catégorie /a/ de l'espagnol dans presque 100% des cas. Quant à la voyelle /ɔ̃/, elle est assimilée à /o/ dans 73.2% chez les colombiens, et un peu plus chez les espagnols, 79.7%. Enfin, /ɑ̃/ a été assimilé à deux catégories de l'espagnol : 75% à /o/, et 24.6% à /a/ par les naïfs colombiens ; de la même manière les espagnols ont catégorisé /ɑ̃/ comme /o/, 58.3% des cas, et comme /a/ dans 40.3%.

Par ailleurs, la catégorisation des voyelles nasales en catégories vocaliques natives par les apprenants montre une assimilation influencée par leur expérience linguistique du français : /ɛ̃/ continue d'être assimilé à /a/, mais il y a un petit nombre de cas en /e/ (5.6%), ainsi qu'une tendance à catégoriser sous l'influence d'une représentation de l'orthographe du français (/ɛ̃/ = /i/, 19.4%). L'assimilation de /ɔ̃/ en /o/ est maintenue et augmentée chez les apprenants. Enfin, l'assimilation de /ɑ̃/ dans deux catégories de l'espagnol, /a/ et /o/, est maintenue, quoiqu'en moindre pourcentage, la double catégorisation est équilibrée : /ɑ̃/ = /a/, 47.6%, et /ɑ̃/ = /o/, 51.2%, chez les apprenants colombiens de niveau avancé ; alors que pour les espagnols de niveau avancé, l'assimilation de /ɑ̃/ à /a/ représente 48.1%, et pour /o/, 37%. Nous observons donc, d'un côté, un affinement phonétique des voyelles nasales du français chez les apprenants, en fonction de l'assimilation en catégories vocaliques de l'espagnol, et d'autre part, une double catégorisation pour /ɑ̃/ : /a/ et /o/ de l'espagnol. L'intégration des résultats d'assimilation et des degrés de confiance est présentée ci-dessous.

### 3.3 Modélisation des performances de discrimination des voyelles nasales

Les résultats d'assimilation devraient permettent de formuler des prédictions sur les performances de discrimination des voyelles nasales par les hispanophones, à la lumière des prédictions du PAM sur la discrimination des contrastes non-natifs. Nous avons trouvé que /ɛ̃/ est assimilé massivement à /a/ et /ɔ̃/ à /o/, tandis que /ɑ̃/ est assimilé soit à /a/ soit à /o/ (/o/ étant dominant). Par conséquent, /ɛ̃/-/ɔ̃/ est un contraste TC (*two category assimilation* : catégories /a/ et /o/) que PAM prédit bien

discriminé. Pour /ɑ̃/-/ɔ̃/, comme /ɔ̃/ est assimilé à /o/ et /ɑ̃/ le plus souvent à /o/, il est le plus souvent SC (*single category assimilation*) ou CG (*category goodness assimilation*) que PAM prédit difficile ou relativement facile. Quant à /ɑ̃/-/ɛ̃/, comme /ɛ̃/ est assimilé à /a/ et /ɑ̃/ le plus souvent à /o/, il est le plus souvent TC, donc facile, mais parfois SC, donc plus difficile, lorsque /ɑ̃/ est assimilé à /a/. Pour nos données d'assimilation, PAM prédirait donc la meilleure discrimination pour /ɛ̃/-/ɔ̃/, suivi par /ɑ̃/-/ɛ̃/, et la moins bonne pour /ɑ̃/-/ɔ̃/. C'est en effet ce que nous trouvons dans un test de discrimination que nous ne rapportons pas en détail ici (Figure 3). Pour affiner les prédictions, il s'agit de savoir dans quelle mesure un contraste est plutôt CG que SC. Typiquement, cette distinction est reflétée par des différences de rating (Best, 1995). Nous avons tenté de quantifier nos prédictions en prenant en compte les différences de rating. Si l'on ignore ces différences, une formule simple dérivant la performance de discrimination des données d'assimilation est $P(\neq/(x,y)) = 1 - \sum_{i=1}^{5} P(ci/x) \times P(ci/y)$, *ci* étant les 5 catégories /a, i, e, o, u/. Le produit de probabilités correspond à la probabilité d'assimiler les deux termes du contraste, x et y, à une même catégorie, donc de ne pas discriminer x et y sur cette catégorie. Notre idée est de pondérer ce terme par un facteur d'autant plus petit que la différence de rating est grande. La probabilité prédite de discriminer x et y, $P(\neq/(x,y))$, est d'autant plus élevée que la différence est plus grande entre les ratings obtenus pour x et y sur la catégorie *ci* pour i de 1 à 5. Pour tester cette idée, nous avons utilisé une pondération très simple : (4 – |rating(x=*ci*) – rating(y=*ci*)|)/4. (Les ratings sont de 1 à 5, donc la plus grande différence est 4.). La Figure 3 montre les données de discrimination observées et prédites par notre "modélisation". Cette modélisation donne un poids aux différences de "category goodness", tout en restant basée sur les différences qualitatives entre les types de catégorisation prévues de PAM : TC versus SC ou CG. Nous nous proposons d'optimiser cette modélisation en utilisant une fonction paramétrable des différences de rating. Pour l'instant, nous ne montrons que la faisabilité de la modélisation.

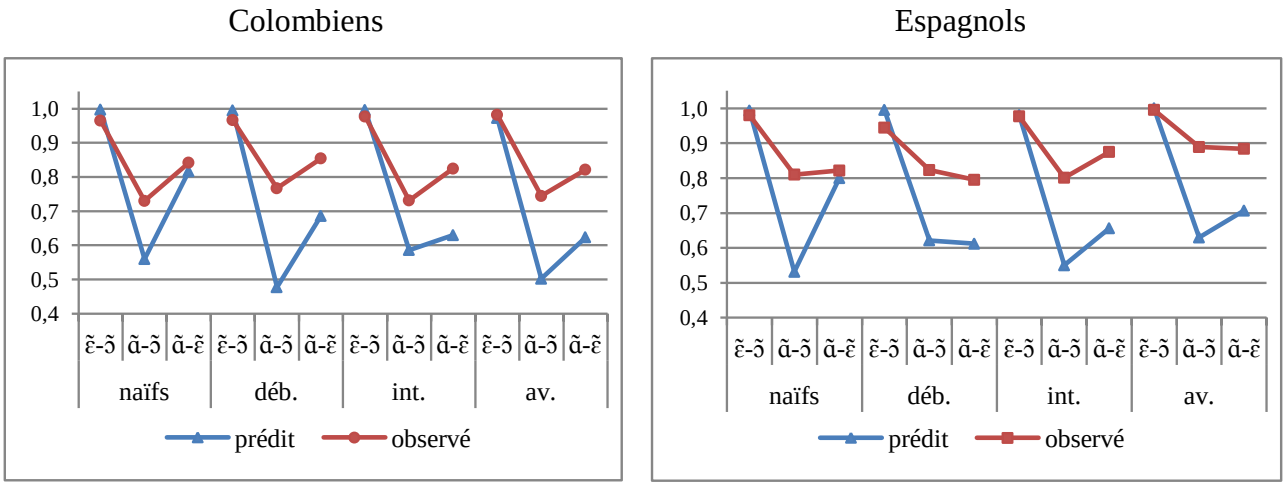



FIGURE 3. Prédiction de la discrimination des contrastes nasales /ɛ̃/-/ɔ̃/, /ɑ̃/-/ɔ̃/, et /ɑ̃/-/ɛ̃/ par les sujets naïfs et les apprenants hispanophones par niveaux de français : Colombiens, à gauche, et Espagnols, à droite ; prédiction en bleu et la discrimination observée en rouge.

Les résultats de discrimination, en rouge, montrent que le contraste /ɛ̃/-/ɔ̃/ est mieux discriminé par rapport aux autres contrastes dans les deux groupes d'hispanophones. Concernant le contraste /ɑ̃/-/ɔ̃/, il est le plus difficile à discriminer par les sujets colombiens, $F(_{2,3237}) = 120.6$, $p < 0.0001$, comme prédit par le modèle quoique de manière plus amplifiée, alors que chez les espagnols les prédictions indiquent aussi une difficulté à discriminer l'opposition /ɑ̃/-/ɔ̃/ notamment, mais les résultats montrent qu'il n'y pas de différence significative de discrimination entre les deux

contrastes /ɑ̃/-/ɔ̃/ et /ɑ̃/-/ɛ̃/ ($p$ = 0.39), même s'il y a une différence significative entre tous les contrastes, $F(_{2,3957})$ = 90.07, $p$ < 0.0001.

# 4. Discussion, conclusion et perspectives

Cette étude cherchait à examiner la perception des voyelles nasales du français par les apprenants hispanophones dans un test d'identification en fonction de la position syllabique des voyelles nasales. Les résultats du test d'identification des voyelles nasales ont montré une confusion entre /ɑ̃/-/ɛ̃/ chez les apprenants colombiens, et une difficulté plus accentuée, en particulier pour l'identification de /ɑ̃/ chez les espagnols. Même si l'identification de /ɔ̃/ est la moins problématique, elle est affectée par la position syllabique, notamment en position initiale absolue. Selon les difficultés observées, nous nous sommes intéressés à étudier comment les sujets naïfs hispanophones du français assimilent les voyelles nasales selon les catégories vocaliques de l'espagnol. Finalement, ces assimilations nous ont permis de formuler une série de prédictions des performances de discrimination des voyelles nasales par les apprenants et les naïfs hispanophones. La catégorisation des voyelles nasales que réalisent les sujets naïfs hispanophones selon les catégories de l'espagnol nous a permis de mieux comprendre les difficultés d'identification, ainsi que de proposer un modèle quantitatif de prédiction de la discrimination des voyelles nasales du français à partir des données de catégorisation et de rating. La double catégorisation de /ɑ̃/ en /a/ et /o/, dont le rating était plus important pour /o/, devient un facteur de difficulté pour les apprenants lorsqu'il est opposé à /ɔ̃/, et c'est ce que nous avons constaté avec les résultats de discrimination chez les colombiens. Quant aux sujets espagnols, les prédictions de discrimination n'ont coïncidé qu'avec les résultats de discrimination de /ɛ̃/-/ɔ̃/ ; cependant leur difficulté avec l'identification de /ɑ̃/ peut être reflétée par leur difficulté à discriminer les contrastes /ɑ̃/-/ɔ̃/ et /ɑ̃/-/ɛ̃/. Nous projetons à l'avenir d'améliorer ce modèle de prédiction des performances de discrimination des voyelles nasales du français par des hispanophones. De plus, un rapport avec des résultats de production est à envisager.

# Remerciements

Mille mercis aux personnes ayant participé à cette étude, aussi bien les sujets naïfs pour le français que les apprenants espagnols et colombiens. Merci aussi à M. Javier Enrique REDONDO, directeur du Département des Langues de la *Pontificia Universidad Javeriana Bogota*, Colombie, et à Mme Inmaculada ILLANES ORTEGA, directrice du *Département de Philologie Française* à l'Université de Séville, Espagne. Ce travail a bénéficié d'une aide du LabEx EFL en 2014 (ANR/CGI). Merci aussi à l'Université Paris Sorbonne-Cité pour la bourse de mobilité sortante 2015.

# Références bibliographiques

BEST, C. T. (1995). "A direct realist perspective on cross-language speech perception," in S*peech Perception and Linguistic Experience: Issues in Cross-language Research*, Ed. W. Strange (Timonium, MD: York Press), 171–204.

BEST, C. T., TYLER, M. D. (2007). "Non native and second-language speech perception: commonalities and complementarities," in M. J. Munro & O.S. Bohn (Eds.), *Second Language Speech Learning: The Role of Language Experience in Speech Perception and Production*, (pp. 13-34). Amsterdam: John Benjamins.

BUSTAMANTE, D. A., AMELOT, A. et PILLOT-LOISEAU, C. (2014). « Étude de la production des voyelles nasales du français chez des apprenantes espagnoles et colombiennes », XXXe édition des Journées d'Études sur la Parole, Le Mans, 23 - 27 juin 2014, 576-580.

DESMEULES-TRUDEL, F. (2013). Perception des voyelles nasales du français québécois : aspects acoustiques et perceptifs. *Mémoire de Maîtrise en Linguistique*, Université de Laval, Québec.

DETEY, S., RACINE, I., KAWAGUCHI, Y., ZAY, F., & BUEHLER, N. (2010). Évaluation des voyelles nasales en français en L2 en production : de la nécessité d'un corpus multitâches. In: Neveu, F., Durand, J., Klingler, T., Prévost S., Muni-Toké V. (éds.). *Actes de CMLF'10* [CD-ROM], ILF, 1289-1301.

DETEY, S. et RACINE, I. (2015). Does perception precede production in the Initial stage of French nasal vowel quality acquisition by Japanese learners? A corpus-based discrimination experiment. *Proceedings of ICPhS2015*, Glasgow, 10-14 August.

ESCUDERO, P. (2005). "Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization," Ph.D. thesis, Utrecht University, Utrecht, Netherlands.

FLEGE, J. E. (1995). ''Second language speech learning: Theory, findings and problems,'' in *Speech Perception and Linguistic Experience: Issues inCross-language Speech Research*, edited by W. Strange (York, Timonium, MD), pp. 233-272.

INCEOGLU, S. (2014). Effect of multimodal training on the perception of French nasal vowels. *Concordia Working Papers in Applied linguistics*, 5, 311-321.

KAKOYIANNI-Doa, F., MONVILLE-BURSTON, M., & ARMOSTIS, S. (2017). "Les nasales /ɛ̃/ et /ɑ̃/ chez les apprenants hellénophones". *Revue du Centre Européen d'Etudes Slaves - La revue | Numéro 6*. [En ligne] Publié en ligne le 06 mars 2017. URL : http://etudesslaves.edel.univ-poitiers.fr/index.php?id=1108 (consulté le 28/11/2017).

KAMIYAMA, T. (2009). *Apprentissage phonétique des voyelles orales du français langue étrangère chez des apprenants japonophones.* Thèse de doctorat. Université de la Sorbonne Nouvelle, Paris.

KUHL, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of SciencesUSA* 97, 11850–11857.

MAGNEN, C., BILLIERES, M. & GAILLARD, P. (2005). Surdité phonologique et catégorisation: perception des voyelles françaises par des hispanophones. *Revue PArole, 33*, 9-30.

NAWAFLEH, A. (2013). *Difficultés de prononciation et de perception de voyelles du français par des apprenants arabophones: apprenants jordaniens.* Presses Académiques Francophones, 535 pages.

TROUBETZKOY, N. S. (1938/1967). *Principes de phonologie*, Paris, Éditions Klincksieck.

# Quel est mon âge d'après ma voix ? Effets de la variété régionale et de la génération.

Nicolas Audibert[1], Cécile Fougeron[1], Fany Barbier[1,2],
Léa Croze[1,2], Camille Lavoine[1,2], Hélène Rance[1,2]
(1) Laboratoire de Phonétique et Phonologie, UMR7018 CNRS/Sorbonne-Nouvelle,
19 rue des Bernardins, 75005 Paris, France
(2) Département Universitaire d'Enseignement et de Formation en Orthophonie,
Sorbonne Université UPMC, 91 Boulevard de l'hôpital, 75013 Paris, France
{nicolas.audibert, cecile.fougeron}@sorbonne-nouvelle.fr

## RESUME

L'âge d'un locuteur peut souvent être estimé à partir de sa voix seule. Nous testons 112 locuteurs francophones âgés de 50 à 89 ans, répartis en 4 variétés régionales (français, belges, suisses, québecois), 2 sexes et 4 décennies. 13 auditeurs français jeunes (22-31) et 13 âgés (70-95) ont estimé l'âge du locuteur comme appartenant à l'une des 4 classes d'âge. L'âge perçu apparaît comme un indicateur assez fiable de l'âge chronologique d'un locuteur, mais avec une tendance à surestimer l'âge des [50-59] ans et sous-estimer l'âge des locuteurs de 70 ans ou plus. L'estimation dépend de l'âge de l'auditeur : les auditeurs âgés surestiment davantage l'âge des locuteurs de la génération précédente [50-69]. L'origine partagée ou non entre auditeur et locuteur affecte aussi les réponses : les auditeurs français estiment plus âgés les locuteurs d'une autre variété régionale.

## ABSTRACT

**Guess my age from my voice. Effect of regional variety and generation.**
The age of a speaker can often be estimated from his/her voice alone. We test 112 speakers of French aged 50 to 89, distributed into 4 regional varieties (French, Belgian, Swiss, Quebecois), 2 sexes and 4 decades. 13 young (22-31) and 13 old (70-95) French listeners estimated the speaker's age as belonging to one of the 4 age classes. Perceived age appears as a fairly reliable cue to the chronological age of a speaker, but with a tendency to overestimate the age of the speakers in the [50-59] group and underestimate the age of speakers from 70 y.o. Estimations depend on listener's age: older listeners overestimate more the age of speakers of the younger generation ([50-69]). The sharing of origin between speaker and listener also affects answers: French listeners rate the speakers of another regional variety as older.

MOTS-CLES : Age, vieillissement, perception, variation régionale, effet de l'auditeur.
KEYWORDS: Aging, perception, dialectal variation, listener effect.

## 1 Introduction

Dans les études concernant les effets de l'âge sur la parole de locuteurs adultes, une question récurrente porte sur la définition de la notion même de vieillissement. A partir de quel âge est-on vieux ? Sur quels critères définir les groupes d'âge comparés dans une étude transversale sur le


10.21437/JEP.2018-70

vieillissement ? Doit-on se référer à l'âge chronologique du locuteur, à l'image renvoyée/perçue, à ses performances cognitives, à des mesures/marqueurs physiologiques, à sa condition physique, etc. ? Concernant ce dernier aspect par exemple, Ramig (1983) a comparé des locuteurs de 3 groupes d'âge chronologique (25-35), (45-55), (65-75), chacun formés de 8 locuteurs en bonne condition et 8 locuteurs en mauvaise condition physique. Que les pauses soient comptabilisées ou non, le débit de parole des locuteurs âgés (groupe 65-75) est plus lent que celui des locuteurs jeunes, quelle que soit la condition physique. Toutefois, ces différences sont plus marquées dans un sens comme dans l'autre chez les locuteurs en mauvaise condition physique. On notera que dans cette étude, le groupe considéré comme « âgé » va de 65 à 75 ans, alors que dans d'autres, comme chez Fletcher et al (2015), les plus jeunes locuteurs vont de 65 à 69 ans et les plus âgés de 85 à 89 ans.

Dans cette littérature sur le vieillissement et ses effets sur la parole, force est de constater que les études sont très peu comparables dans leur définition des « locuteurs jeunes » par rapport aux « locuteurs âgés ». Certains se basent sur une définition « occidentale » (cf Pierce et al. 2013) de la personne âgée, avec une césure liée au changement d'activité vers 65 ans, d'autres prennent en compte des changements physiologiques (ex. changement hormonaux chez la femme à partir de 50 ans, réduction des fibres nerveuses à partir de 60), des changements neurologiques affectant le contrôle moteur ou les fonctions cognitives. Outre ces critères hétérogènes, la variabilité dans la définition des groupes d'âge est accrue par la nécessité dans de nombreuses études de constituer des groupes de taille homogène en fonction des locuteurs disponibles. Quels que soient les critères de groupement, on sait qu'une forte variabilité inter-individuelle est à attendre d'un découpage en âge chronologique et que l'hétérogénéité des groupes augmente pour les locuteurs âgés. En effet, l'âge accentue les différences entre les individus au niveau biologique (Woodruff & Birren 1975), mais aussi au niveau de la prise de médication ou de l'isolement social qui peuvent aussi affecter la parole. Afin de pallier cette hétérogénéité dans les groupes, plusieurs études préfèrent se référer à l'âge perçu du locuteur qu'à son âge chronologique. Pour autant, comme toutes mesures issues d'un traitement perceptif subjectif, mesurer un âge perçu n'est pas exempt de problèmes.

En effet, même si tout un chacun peut prétendre avoir déjà estimé l'âge d'un locuteur à partir de sa voix, au téléphone par exemple, la littérature montre que cette estimation peut être influencée par plusieurs facteurs. Plusieurs études ont montré que les auditeurs pouvaient estimer l'âge d'un individu à partir de sa voix de façon relativement précise. De bonnes corrélations entre âge perçu et âge chronologique ont été trouvées dans la littérature et dans plusieurs langues (corrélations comprises entre r=.66 et r=.91 pour l'anglais américain, l'allemand, l'italien et le japonais, voir Hunter & Ferguson, 2016 pour une revue). Pourtant, malgré ces corrélations élevées, les études montrent une tendance à la sous-estimation de l'âge perçu pour les locuteurs les plus âgés (Huntley et al., 1987, Kido & Kasuya, 2004, Hunter & Ferguson 2016). La précision de l'estimation semble également dépendre de l'âge de l'auditeur, ou plutôt du décalage de génération entre l'auditeur et le locuteur. Pour Hollien & Tolhurst (1978), les auditeurs n'arrivent à évaluer l'âge que de locuteurs qui leurs sont proches en âge. Pour d'autres, ce n'est pas une question de « familiarité avec la génération », mais une capacité d'estimation qui est moins bonne chez les auditeurs âgés par rapport au plus jeunes (Kreiman & Papçun, 1985 ; Lindville & Korabic 1986 ; Goy, Pichora-Fuller & van Lieshout, 2016). Enfin, il semblerait que le traitement des indices de l'âge dans la voix et la parole d'un individu ne relève pas uniquement d'un traitement acoustique. En effet, d'après Nagao (2006), il est plus facile d'estimer l'âge d'un locuteur de sa propre langue que dans une autre langue, tandis que Braun et Cerrato (1999) ne trouvent pas de bénéfice à partager la même langue que le locuteur.

Dans cette étude, nous chercherons à mieux comprendre les processus en jeu dans la définition d'un âge perçu, avec comme objectif à plus long terme de pouvoir analyser les caractéristiques

acoustiques de la parole de personnes âgées en fonction de leur âge chronologique vs. perçu. Pour l'heure, nous chercherons a savoir comment l'estimation de l'âge du locuteur varie en fonction de son âge chronologique et en fonction de caractéristiques partagées ou non entre le locuteur et l'auditeur qui le juge : la génération et la variété régionale du français.

# 2 Méthode

## 2.1 Locuteurs et prétraitement des stimuli

112 locuteurs francophones âgés de 50 à 89 ans ont été sélectionnés dans la base MonPaGe_HA (Fougeron et al., 2018). Les locuteurs se répartissent en quatre variétés régionales : 32 français (Île-de-France), 32 belges (Mons), 24 suisses (Genève), 24 québécois (Montréal). Au sein de chaque groupe régional, la distribution est équilibrée entre les deux sexes et quatre classes d'âge (50-59, 60-69, 70-79, 80-89), dont les caractéristiques sont données dans la Table 1. La production en lecture de la phrase entièrement voisée « Anne-Marie et moi allons à la mer » (16 phonèmes, 10 syllabes) a été choisie car elle présentait un bon compromis entre le besoin d'avoir suffisamment de matière pour un jugement perceptif et les contraintes de durée de l'expérience. Les enregistrements ont été normalisés en intensité, et recoupés pour d'éliminer les éventuels bruits extérieurs audibles présents au début et à la fin des enregistrements. Les parties coupées ont été remplacées par un silence de 200 ms, avec un lissage sur 5 ms pour éviter les artefacts de lecture audio.

| Groupe | Sexe | Français | Belges | Suisses | Québécois |
|---|---|---|---|---|---|
| 50-59 ans | F | 53 (1.2) | 53.5 (3) | 53.7 (2.1) | 52 (1.6) |
| | H | 51 (1.2) | 52.8 (1.1) | 52.3 (1.2) | 54 (2.9) |
| 60-69 ans | F | 64.3 (2.8) | 64.5 (1.7) | 64.7 (3.1) | 63.7 (1.7) |
| | H | 64 (2.4) | 64.5 (1.8) | 64.7 (0.9) | 63.3 (2.9) |
| 70-79 ans | F | 76.8 (2.3) | 73.3 (3.1) | 75.3 (1.7) | 71.7 (1.2) |
| | H | 75 (2.1) | 77 (1.4) | 73 (3.6) | 74 (2.2) |
| 80-89 ans | F | 82.5 (1.8) | 85.5 (2.5) | 83 (2.4) | 83.3 (2.1) |
| | H | 84 (2.7) | 83.5 (1.1) | 85.7 (3.3) | 81 (0.8) |

TABLE 1 : Moyenne (écart-type) de l'âge des 112 locuteurs dans chaque sous-groupe. Chaque case correspond à 4 locuteurs pour français et belges, 3 pour suisses et québécois.

## 2.2 Test de perception et auditeurs

La tâche principale du test de perception était d'identifier l'âge du locuteur comme appartenant à l'une des quatre décades possibles : 50-59 ans ; 60-69 ; 70-79 ; 80-89. Les choix de réponses proposés aux auditeurs sont donc distribués de façon équilibrée relativement à l'âge chronologique des locuteurs, ce qui correspond à l'approche majoritairement adoptée dans la littérature pour une évaluation catégorielle. Il était également demandé d'évaluer le sexe du locuteur, toutefois les résultats correspondants ne sont pas traités ici. Afin d'évaluer la consistance intra-auditeur, les productions de 16 locuteurs (8 français, 8 belges) ont été incluses deux fois dans le test, donnant un total de 128 stimuli à traiter dans la version complète du test. En raison de leur fatigabilité plus importante, une version réduite du test sans les 48 stimuli suisses et québécois a été mise en place à

destination d'auditeurs âgés, pour un total de 80 stimuli présentés. Un ordre aléatoire différent était proposé à chaque auditeur, la randomisation étant effectuée avec la contrainte de ne pas présenter deux fois consécutives les stimuli inclus en double pour l'évaluation de la consistance intra-auditeur. Un questionnaire a été soumis à l'ensemble des auditeurs avec des questions portant sur la ou les professions ayant été exercées, les langues parlées et les lieux de vie afin de contrôler le degré d'exposition à différents accents. Une évaluation de la qualité de l'audition pour chacune des deux oreilles a également été incluse, ainsi qu'une question sur le port éventuel de verres correcteurs.

La version complète du test, d'une durée totale d'environ 35 minutes en incluant la réponse au questionnaire, a été passée par *13 femmes étudiantes en 4ème année d'orthophonie âgées de 22 à 31 ans (25.1 en moyenne).* Pour ce groupe d'auditeurs, une question spécifique portant sur le degré d'exposition aux troubles de la voix ou de la parole a été ajoutée et fera l'objet d'une analyse ultérieure des différences inter-auditeurs. La version réduite du test a été passée par *13 personnes (11 femmes, 2 hommes) âgées de 70 à 95 ans (86.6 en moyenne)* recrutées au sein d'un EPHAD de banlieue parisienne. La durée totale était d'environ 50 minutes en incluant la réponse au questionnaire et la passation du test d'évaluation des fonctions cognitives MMSE (Mini Mental State Examination). Ces auditeurs âgés ont été sélectionnés avec l'aide du médecin gériatre assurant leur suivi afin de n'inclure que des personnes considérées non-dépendantes et ne souffrant pas de déficits cognitifs avérés, sans troubles visuels ni auditifs majeurs.

# 3 Résultats

La performance des auditeurs pour la réalisation de la tâche de catégorisation de l'âge des locuteurs est évaluée par les mesures suivantes, globalement ou par sous-groupe de locuteurs ou d'auditeurs :

1. Taux de réponses « correctes », c'est-à-dire pour lesquelles la classe d'âge choisie par l'auditeur est celle correspondant à l'âge chronologique du locuteur ;
2. Corrélation de Spearman entre l'âge chronologique du locuteur et la classe d'âge choisie par l'auditeur, considérée comme une variable ordinale ;
3. Nombre de décennies d'écart entre la classe d'âge choisie par l'auditeur et celle correspondant à l'âge chronologique du locuteur, converti en nombre d'années pour plus de clarté.

## 3.1 Performances globales : une tâche complexe

Tous locuteurs et auditeurs confondus, on observe que le taux de réponses « correctes » n'est que de 38%, et la corrélation entre âge du locuteur et classe d'âge choisie de r=.45. Bien que ce résultat soit significativement meilleur que le hasard ($\chi^2(1)=98.08$, $p<2.10^{16}$), permettant ainsi de rejeter l'hypothèse d'une absence d'effet de l'âge sur les jugements, l'âge perçu des locuteurs ne correspond donc pas exactement à leur classe d'âge chronologique. L'analyse de l'écart entre classe d'âge choisie et classe d'âge de référence permet toutefois de nuancer ce tableau : les confusions concernent en effet surtout les décennies adjacentes, avec un écart moyen en valeur absolue de 8.3 ans. Seules 18% de l'ensemble des réponses sont données avec plus d'une classe d'âge d'écart.

La tâche de catégorisation de l'âge des locuteurs s'avère complexe, comme l'indique l'analyse de la fiabilité intra-auditeur à partir des réponses aux 16 stimuli inclus en double : globalement, ces stimuli ne sont évalués à l'identique lors des deux présentations que dans 58% des cas. Là encore, l'analyse de l'amplitude de l'écart entre les deux présentations permet de nuancer : l'écart moyen en valeur absolue n'est que de 4.8 ans, avec seulement 6% des réponses pour lesquelles on observe plus d'une classe d'âge d'écart entre les deux présentations.

## 3.2 Effet de l'âge des auditeurs sur l'estimation de l'âge des français et belges

Les performances des auditeurs jeunes (ci-après AJ) et âgés (ci-après AA) sont comparées sur l'ensemble des stimuli communs aux deux versions du test (soit 64 locuteurs français et belges). Les performances sont globalement différentes entre auditeurs jeunes et les auditeurs âgés. En effet le groupe AJ obtient un taux de réponses « correctes » de 41%, un écart moyen de 7.6 ans et une corrélation entre âge chronologique du locuteur et catégorie de réponse de r=.55, tandis que le groupe AA obtient un taux de réponses « correctes » de 32%, un écart moyen de 9.8 ans et une corrélation de r=.29. Ainsi pour les auditeurs jeunes, l'âge perçu des locuteurs se rapproche plus de leur âge chronologique. On peut en revanche noter que la consistance intra-auditeur moyenne est meilleure pour AA que pour AJ, tant en termes de taux de réponses « correctes » (63% vs. 53%) que d'années d'écart (4.3 vs. 5.3).

L'effet de différents facteur sur l'écart entre classe d'âge choisie et classe d'âge du locuteur est évalué au moyen d'un modèle linéaire mixte, à l'aide du package de R lme4 (Bates et al., 2015). L'utilisation d'un modèle paramétrique supposé s'appliquer à une variable continue pourrait être problématique avec cette variable qui ne peut prendre que 7 valeurs entières comprises entre -3 et +3. Toutefois plusieurs études (par ex. Norman, 2010) ont montré que des variables ordinales telles que des échelles de Likert pouvaient être analysés au moyen de modèles paramétriques sans pour autant accroître le risque d'erreur de type I ou II (respectivement rejet ou acceptation à tort de H0).

Le groupe d'âge des auditeurs, celui des locuteurs et la variété régionale des locuteurs ont été inclus dans le modèle en tant qu'effets fixes, l'interaction entre âge des auditeurs et âge des locuteurs étant également prise en compte. Les variations liées aux différences inter-locuteurs et inter-auditeurs ont été modélisées par une structure aléatoire incluant des pentes aléatoires pour les différences entre locuteurs en fonction du groupe d'âge des auditeurs, ainsi que pour les différences entre auditeurs en fonction du sexe et du groupe d'âge des locuteurs. Bien qu'on ne puisse supposer a priori que tous les auditeurs traitent de façon similaire les différences entre variétés régionales des locuteurs, la non-convergence du modèle avec la pente aléatoire correspondante nous a conduit à ne modéliser cet aspect de la variabilité inter-auditeur que par une ordonnée à l'origine (*intercept*) aléatoire. Le modèle retenu, pour lequel l'inspection des résidus ne révèle pas de déviation majeure aux conditions de normalité et d'homoscédasticité, est donc (loc désignant le locuteur et aud l'auditeur) :

ecartAge ~ groupeAgeAud * groupeAgeLoc + varieteRegionaleLoc +
(1+ groupeAgeAud | loc) + (1+sexeLoc | aud) + (1+ groupeAgeLoc | aud)

Les effets principaux et l'interaction sont illustrés par la Figure 1. La significativité de l'effet des facteurs fixes a été testée en comparant le modèle sans interaction au même modèle sans le facteur considéré dans un test de vraisemblance, et l'effet de l'interaction en comparant le modèle avec vs. sans interaction. Ces comparaisons révèlent un effet significatif de l'ensemble des facteurs fixes ($\chi^2(16)=73.60$, $p=7.10^{16}$ pour le groupe d'âge des locuteurs, $\chi^2(1)=12.63$, p=.0003 pour la variété régionale), le groupe d'âge des auditeurs excepté ($\chi^2(1)=2.45$, p=.117). En revanche, l'interaction entre le groupe d'âge des auditeurs et celui des locuteurs est significative ($\chi^2(3)=14.07$, p=0.003).

Pour cette interaction, la significativité du contraste entre AJ et AA a été estimée pour chaque groupe d'âge de locuteurs. Tandis que cette différence est significative pour les groupes d'âge 50-59 (t(45.15)=2.76, p=.008) et 60-69 (t(42.36)=2.17, p=.036), elle ne l'est ni pour le groupe d'âge 70-79 (t(41.28)=0.21, p=.083) ni pour le groupe 80-89 (t(38.31)=-1.34, p=.189). Ces contrastes nous indiquent donc que les auditeurs âgés surestiment plus l'âge des locuteurs les plus jeunes que ne le

font les auditeurs jeunes, tandis que l'évaluation de l'âge des locuteurs les plus âgés est similaire entre groupes d'auditeurs. Quant à eux, les contrastes entre groupes d'âge de locuteurs au sein de chaque groupe d'âge d'auditeur sont tous significatifs ($p<0.037$ pour toutes les comparaisons), à l'exception du contraste entre locuteurs de 70-79 ans vs. 80-89 ans évalués par le groupe d'auditeurs jeunes ($t(68.79)=2.13$, $p=.155$).

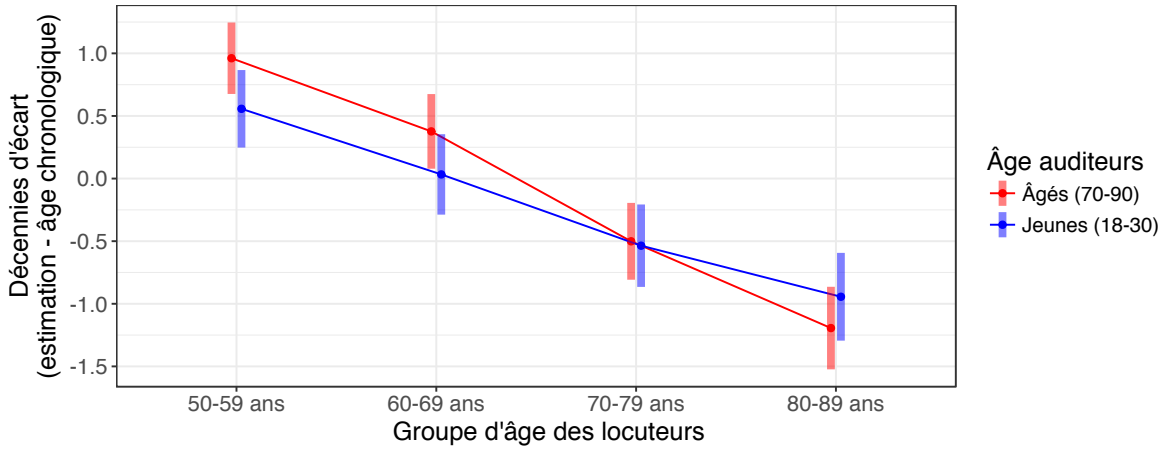



FIGURE 2: Effet du groupe d'âge des locuteurs, du groupe d'âge des auditeurs et de l'interaction, valeurs prédites par le modèle. Les barres verticales représentent les intervalles de confiance.

## 3.3 Effet de la variété régionale des locuteurs sur le jugement des auditeurs jeunes

Sur l'ensemble des stimuli évalués par les auditeurs jeunes, c'est pour les locuteurs suisses que les performances sont les meilleures (47% de réponses « correctes », écart moyen de 6.6 ans en valeur absolue et corrélation de r=.63). En revanche les niveaux de performance dans les trois autres groupes de locuteurs sont très proches les uns des autres (taux de réponses « correctes » entre 39% et 40%, écart moyen entre 7.5 et 7.8 ans), hormis pour les corrélations entre âge du locuteur et catégorie de réponse en raison de différences dans la direction des décalages : r=.58 pour les français, r=.52 pour les belges, et r=.49 pour les québécois. La consistance intra-auditeur, évaluée sur les français et belges, est quant à elle plus élevée pour les locuteurs français (56% de réponses « correctes », 4.9 ans d'écart moyen) que belges (51% de réponses « correctes », 5.7 ans d'écart).

De même que précédemment, l'effet de la variante régionale compte tenu des groupes d'âge des locuteurs a été analysé au moyen d'un modèle linéaire mixte (effets illustrés par la Figure 3) avec l'écart entre classe d'âge choisie et classe d'âge du locuteur comme variable dépendante : ecartAge ~ groupeAgeLoc * varieteRegionaleLoc + (1+sexeLoc | aud) + (1+ groupeAgeLoc | aud)

L'inspection visuelle des résidus ne révèle pas de déviation majeure aux conditions de normalité et d'homoscédasticité. Les tests de vraisemblance révèlent un effet significatif de la variété régionale ($\chi^2(3)=62.32$, $p=2.10^{-13}$) et du groupe d'âge des locuteurs ($\chi^2(3)=48.24$, $p=2.10^{-10}$), mais pas de l'interaction entre ces deux facteurs ($\chi^2(9)=12.35$, $p=.194$). Les contrastes entre variétés régionales nous indiquent que l'âge des locuteurs français est estimé comme significativement moins élevé que celui des locuteurs d'autres variétés régionales ($p<.0001$ pour les trois comparaisons, $p>0.55$ pour les comparaisons entre autres variétés du français). De plus, les mêmes contrastes entre variétés régionales au sein de chaque groupe d'âge révèlent que cette différence entre locuteurs français et autres variétés régionales se neutralise pour les locuteurs de 80 à 89 ans (tous $p>209$).

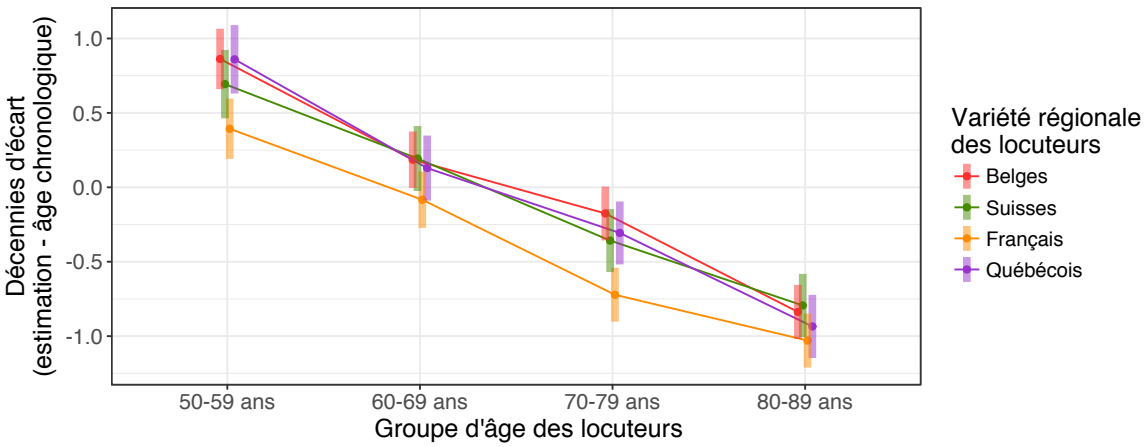



FIGURE 3: Effet du groupe d'âge, de la variété régionale des locuteurs et de l'interaction, valeurs prédites par le modèle. Les barres verticales représentent les intervalles de confiance.

# 4 Discussion et conclusion

Le premier résultat qui ressort de cette étude montre que l'âge perçu de nos locuteurs couvrant 4 décennies successives (plutôt que des groupes très éloignés en âge) correspond relativement bien à leur âge chronologique, à environ 8 ans près. Toutefois, on observe une tendance à surestimer l'âge du premier groupe ([50-59] ans) et à sous-estimer l'âge des locuteurs des deux derniers groupes. En d'autre termes, c'est dans le groupe [60-69] ans que les locuteurs « sonnent » le plus proche de leur âge chronologique, alors que les [50-59] paraissent plus vieux qu'ils ne le sont et les plus de 70 ans, plus jeunes qu'ils ne le sont. Ce résultat est toutefois à pondérer par le fait que dans ce type de tâche les auditeurs ont tendance à centraliser les réponses vers le milieu de l'échelle et que les faibles taux d'accord intra-juge pour les doubles jugements en test-retest montrent que la tâche demandée aux auditeurs est relativement complexe. Il serait intéressant de comparer sur les mêmes locuteurs, les résultats de cette tâche de catégorisation en quatre classes d'âge à ceux d'une tâche, que nous n'avons pas retenue car nous semblant a priori plus complexe, dans laquelle on demande à l'auditeur une estimation de l'âge exact du locuteur. Pour autant, la tendance à la surestimation de l'âge des locuteurs plus jeunes et la sous-estimation des locuteurs les plus âgés que nous observons se retrouve aussi dans la littérature. Par exemple, Hunter & Ferguson (2017) observent que les jugements de l'âge d'un même locuteur enregistré longitudinalement sur 50 ans tendent à la surestimation de son âge dans la cinquantaine et une sous-estimation de son âge après 70 ans.

Le second résultat important concerne la différence dans les jugements en fonction de l'âge des auditeurs. L'âge perçu d'un locuteur est plus proche de son âge chronologique lorsque cet âge perçu est estimé par des auditeurs jeunes par rapport à des auditeurs âgés. L'analyse des interactions montre par ailleurs que c'est particulièrement sur les locuteurs de moins de 70 ans (50-59 et 60-69) que l'effet de l'âge de l'auditeur est important. Contrairement à ce qu'on aurait pu attendre, ce n'est pas pour l'estimation de l'âge des locuteurs les plus âgés que les juges diffèrent : au-delà de 70 ans les locuteurs sont perçu comme plus jeunes –en d'autres termes leur âge est sous-estimé– par les auditeurs jeunes et âgés. Au contraire, ce sont les locuteurs de 50 à 69 ans qui voient leurs âges davantage surestimés (et sont donc perçus plus âgés que leur âge chronologique) lorsqu'ils sont jugés par des auditeurs âgés que par des auditeurs jeunes. Ces résultats trouvent écho dans la littérature où une différence de performance entre auditeurs âgés et auditeurs jeunes a été relevée (e.g. Lindville & Korabic 1986; Goy, Pichora-Fuller & van Lieshout, 2016). Cette différence ne semble pas liée à une difficulté particulière des auditeurs âgés pour cette tâche comme le suggèrent

Kreiman & Papçun (1985). En effet dans notre expérience, les auditeurs âgés semblent même globalement meilleurs en termes d'accord intra-juge que les auditeurs jeunes. Huntley, Hollien, & Shipp (1987) expliquent les différences entre groupe de juges, non pas par une différence d'âge en soi (dans leur étude les personnes âgés et les adolescents montrent des performances différentes de celles des adultes d'âge moyen), mais par une différence de familiarisation et d'expérience avec certaines voix, provoquant un bénéfice dans le jugement de ses pairs. Dans notre étude, cela signifierait que les auditeurs âgés qui ont plus de 70 ans surestimeraient davantage l'âge des locuteurs n'appartenant pas à la même génération ([50-69]). Dans ce cas que dire des auditeurs jeunes (22-31 ans) qui n'appartiennent à aucune des générations représentées parmi les locuteurs à tester ? Il semblerait que les locuteurs les moins jeunes au sein de ce groupes de jeunes aient de meilleures performances que les plus jeunes, comme le montre la corrélation dans le groupe AJ entre l'âge de l'auditeur et les performances (r=.67 pour le taux de bonnes réponses, r=-.68 pour le nombre de décennies d'écart). Pour autant, l'explication des estimations globalement plus justes/proches de l'âge chronologique de la part du groupe d'auditeurs jeunes pourrait tout simplement résider dans le fait que ce groupe, constitué d'orthophonistes en fin de formation, est probablement plus « expert » en écoute de voix à des fins d'évaluation que le groupe de locuteurs âgés. La comparaison entre le bénéfice acquis par l'expérience et la familiarisation par rapport à celui acquis par l'expertise dans les capacités analytiques d'écoute fera l'objet d'une étude future comparant les évaluations d'auditeurs experts naïfs à celles d'auditeurs naïfs appariés en âge.

Enfin, nous avons pu également montrer que l'origine partagée ou non entre auditeur et locuteur affecte aussi les réponses. Dans cette étude, tous les auditeurs sont français et estiment plus âgés les locuteurs d'une autre variété régionale que la leur. Ici encore, il serait possible d'expliquer ce résultat en termes de familiarisation avec les voix à juger, à l'image des meilleures performances obtenues pour les auditeurs connaissant le locuteur évalué (Hunter & Ferguson ,2017). Toutefois, la tendance à surestimer l'âge des locuteurs présentant un éventuel accent régional est aussi à mettre en rapport avec les résultats de Stölten & Engstrand (2003) qui ont montré une corrélation entre la perception de l'âge et la perception de la force d'un accent dialectal en suédois, cet accent étant perçu comme plus fort chez les sujets âgés et inversement. Drager (2010) a également montré que la perception d'un auditeur, en l'occurrence l'identification des voyelles, dépend des caractéristiques sociales (origine, âge) qu'il attribue au locuteur. Afin d'approfondir ces hypothèses, il sera nécessaire d'examiner les réponses d'auditeurs belges, suisses et québécois sur les mêmes données.

# Remerciements

Ce travail est soutenu en partie par l'ANR VoxCrim (ANR-17-CE39-0016) et le Labex EFL (ANR-10-LABX-0083).

# Références

BATES, D., MAECHLER, M., BOLKER, B., WALKER S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.

BRAUN, A., CERRATO, L. (1999). Estimating speaker age across languages. Actes de *ICPhS 1999*, 1369-1372.

CERRATO, L., FALCONE, M., PAOLONI, A. (2000). Subjective age estimation of telephonic voices. *Speech Communication*, 31(2-3), 107-112.

DRAGER, K. (2011). Speaker age and vowel perception. *Language and Speech*, 54(1), 99-121.

FLETCHER, A. R., MCAULIFFE, M. J., LANSFORD, K. L., LISS, J. M. (2015). The relationship between speech segment duration and vowel centralization in a group of older speakers. *The Journal of the Acoustical Society of America*, 138(4), 2132-2139.

FOUGERON C., DELVAUX V., MÉNARD L., LAGANARO M. (2018) The MonPaGe_HA Database for the Documentation of Spoken French Throughout Adulthood, Actes de *LREC 2018*.

GOY H, PICHORA-FULLER K.M., VAN LIESHOUT P. (2016) Effects of age on speech and voice quality ratings. *The Journal of the Acoustical Society of America,* 139, 1648-1659.

KREIMAN, J., PAPÇUN, G. (1985). Voice discrimination by two listener populations. *The Journal of the Acoustical Society of Americ*a, 77, S9.

HOLLIEN, H., TOLHURST, G. (1978). The aging voice. In *Transcripts of the 7th symposium care of the professional voice, II: Life span changes in the human voice,* 67-73. New York: Voice Foundation.

HUNTER, E. J., FERGUSON, S. H., NEWMAN, C. A. (2016). Listener estimations of talker age: A meta-analysis of the literature. *Logopedics Phoniatrics Vocology*, 41(3), 101-105.

HUNTLEY, R., HOLLIEN, H., SHIPP, T. (1987). Influences of listener characteristics on perceived age estimations. *Journal of Voice*, 1(1), 49-52.

LINVILLE, S. E., KORABIC, E. W. (1986). Elderly listeners' estimates of vocal age in adult females. *The Journal of the Acoustical Society of America*, 80(2), 692-694.

NAGAO, K. (2006). *Cross-language study of age perception*. Doctoral dissertation, Indiana University, Bloomington.

NORMAN, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, 15(5), 625-632.

PIERCE, J. E., COTTON, S., PERRY, A. (2013). Alternating and sequential motion rates in older adults. *International journal of language & communication disorders*, 48(3), 257-264.

PTACEK, P. H., SANDER, E. K. (1966). Age recognition from voice. *Journal of Speech, Language, and Hearing Research*, 9(2), 273-277.

RAMIG, L. A. (1983). Effects of physiological aging on speaking and reading rates. *Journal of communication disorders*, 16(3), 217-226.

SHIPP, T., HOLLIEN, H. (1969). Perception of the aging male voice. *Journal of Speech, Language, and Hearing Research*, 12(4), 703-710.

STÖLTEN, K., ENGSTRAND, O. (2003). Effects of perceived age on perceived dialect strength: A listening test using manipulations of speaking rate and F0. *Phonum*, 9, 29-32.

WOODRUFF, D. S., BIRREN, J. E. (1975). *Aging: Scientific perspectives and social issues*. New York : D. van Nostrand.

# Représentation et Estimation de la Force de Voix à partir du Spectre Moyen à Long Terme

Jean-Sylvain Liénard
LIMSI, 91405 Orsay cedex, France
jean-sylvain.lienard@limsi.fr

## RESUME

La présente étude vise à retrouver par le calcul le niveau sonore émis par un locuteur, à partir de la seule enveloppe du spectre à long terme. Les données utilisées consistent en un ensemble de Spectres Moyens à Long Terme en tiers d'octave, étalonnés en niveau sonore et comportant une grande variabilité en fonction du genre du locuteur, de son âge et du degré d'effort vocal requis. La représentation visuelle des spectres montre qu'il est plus cohérent de les regrouper en fonction du niveau émis qu'en fonction du degré d'effort vocal requis. Une procédure de comparaison est appliquée à l'ensemble des spectres, après normalisation à une valeur commune, arbitraire, de leur niveau sonore. Les résultats indiquent que la forme du spectre est suffisante pour retrouver le niveau sonore émis, avec une marge d'erreur statistique inférieure à 5 dB.

## ABSTRACT

**Representing and Recovering Voice Strength from the Long Term Average Spectrum**

The goal of the study is to recover the Sound Pressure Level emitted by a speaker, from the single long term spectrum envelope. The data consists of a set of 1/3rd octave Long Term Average Spectra, calibrated in sound level and exhibiting a large variability according to the speaker's gender, age and requested vocal effort degree. The visual representation of the spectra shows that it is more coherent to group them according to the emitted sound level than from the requested vocal effort degree. A comparison procedure is then applied to the data, after normalization of the spectra to a common, arbitrary value of their sound level. The results indicate that the single spectral envelope is sufficient to recover the emitted sound level, within a statistical margin of error smaller than 5 dB.

MOTS-CLES : Effort Vocal, Force de Voix, Spectre Moyen à Long Terme.
KEYWORDS: Vocal Effort, Voice Strength, Long Term Average Spectrum.

## 1 Introduction

L'Effort vocal (EV) joue un rôle essentiel dans l'interaction orale: le parleur ajuste l'intensité de sa voix, selon la situation, de façon à se faire bien comprendre par son interlocuteur. Ce faisant il modifie notablement les structures spectro-temporelles du signal qu'il émet. Ces modifications sont reconnues par l'interlocuteur, qui peut ainsi juger de l'intensité émise, indépendamment du


10.21437/JEP.2018-71

niveau sonore parvenant à son oreille. Si l'on se place hors du contexte de l'interaction orale, ces modifications apparaissent comme une variabilité indésirable, qui complique la recherche des structures phonétiques du signal de parole ainsi que son traitement automatique.

Les travaux menés antérieurement dans la recherche des effets de l'EV sur les structures de la parole (HANSON, 1997; HUBER et al., 1999; LIENARD and DI BENEDETTO, 1999; TRAUNMULLER and ERIKSSON, 2000) ou dans la perspective de retrouver le niveau émis (LIENARD and BARRAS, 2013; LIENARD, 2014) portent sur des voyelles isolées ou de très courtes phrases et ne s'appliquent guère à la voix "conversationnelle" (de très faible à très forte) utilisée majoritairement dans les situations ordinaires de l'interaction orale. Ces travaux s'appliquent encore moins aux extrêmes que sont les voix criées ou chuchotées, qui correspondent à des situations exceptionnelles dans lesquelles on doit accepter une perte d'intelligibilité.

En voix conversationnelle les déformations spectro-temporelles induites par les variations d'effort vocal peuvent être considérées comme des variations de timbre. Le timbre individuel du locuteur, ou celui qu'il donne à sa voix pour faire passer telle ou telle émotion ou expression stylistique, se traduit en partie par des variations de force de voix (RILLIARD et al., 2018).

L'étude vise à mettre en évidence les modifications spectrales causées par l'effort vocal et à en déduire une estimation de l'intensité fournie par le locuteur. Cette démarche se heurte à deux difficultés. La première est que la notion d'EV elle-même est mal définie, avec des qualificatifs tels que voix modale, normale, faible, forte, criée, confidentielle, sourde, feutrée, stridente, etc., qui relèvent tout autant de la dimension de timbre que de celle d'intensité. La seconde tient à la mesure de l'intensité de la voix, nécessaire pour une approche objective du problème.

L'intensité est souvent négligée en phonétique, comme en traitement automatique de la parole. Au mieux elle est utilisée en valeur relative et caractérisée par sa variation au long d'une même séquence orale. Nous nous intéressons ici à l'intensité sonore dans l'absolu, évaluée sur une durée de plusieurs secondes afin de moyenner les variations syllabiques et prosodiques. La distinction entre intensité émise (par le parleur) et intensité reçue (par l'auditeur ou par le microphone) est essentielle. Pour réduire le risque d'ambiguïté ainsi que l'imprécision de la notion d'Effort Vocal nous emploierons le terme de Force de Voix (FDV) pour désigner l'intensité moyenne émise par seconde de signal (niveau équivalent $L_{EQ}$, en décibels).

Il n'existe pas aujourd'hui de base de données publique de grande dimension qui permette d'associer à une séquence de parole une mesure fiable de la FDV. Pourtant une telle base de données a été ralisée en 1977 par Pearsons et al. dans le but d'élaborer des normes d'intelligibilité dans le bruit (PEARSONS et al., 1977). Les enregistrements sonores ont été égarés par la suite, mais les relevés de mesure (Spectres Moyens à Long Terme, SMLT) ont été retrouvés, numérisés et mis à disposition de la recherche par Anthony Nash (NASH, 2014). C'est sur ces relevés que porte la présente étude, qui comporte deux parties principales. La première consiste en une analyse qualitative des données, à partir de leur représentation graphique. La seconde vise à retrouver la FDV par le calcul, à partir de la forme du spectre après recalage de tous les SMLT à une valeur commune arbitraire de leur FDV.

# 2 Représentation de l'Effort Vocal

Dans cette partie les données sont représentées graphiquement dans le but de faire apparaître qualitativement les relations entre le SMLT et l'EV. Ces relations, relativement floues si l'on décrit l'EV par la consigne d'effort vocal donnée aux locuteurs, s'avèrent beaucoup plus nettes si l'EV est représenté par la mesure effective de la FDV.

## 2.1 Les données de Pearsons et al.

L'objectif de l'équipe de Pearsons était de définir le niveau de bruit maximum tolérable dans des lieux publics ou privés tels que trains, avions, hôpitaux, écoles, appartements, sans compromettre l'intelligibilité de la parole. Des mesures extensives et soignées du niveau sonore ont été effectuées dans ces lieux. De plus, un corpus de parole a été enregistré dans une chambre anéchoïque en conditions contrôlées (à 1 m dans l'axe de la bouche du sujet, avec le microphone du sonomètre lui-même). Le matériau de parole était une phrase sans signification, phonétiquement équilibrée ("Joe took father's shoe bench out; she was waiting at my lawn"), traditionnellement utilisée dans les tests de qualité des systèmes téléphoniques. Il était demandé aux 97 locuteurs non-professionnels (48 hommes, 37 femmes, 12 enfants de moins de 13 ans) de prononcer cette phrase de manière répétitive pendant au moins 10 secondes, selon 4 consignes vocales: "normal", "raised" (appuyé), "loud" (fort), "shout" (crié). De plus une conversation informelle entre le sujet et l'opérateur distant de 1 m s'ajoutait aux précédents enregistrements et recevait le qualificatif de "casual" (détendu, décontracté). Il s'agissait dans ce cas de voix relativement faible et le contenu phonétique n'était pas spécifié. Le nombre total d'enregistrements s'élevait à 482.

Les sons enregistrés étaient traités ensuite par un analyseur (banc de 24 filtres en tiers d'octave, dont les fréquences centrales allaient de 50 Hz à 10 kHz). L'énergie par canal, sommée tout au long de la séquence, était conservée et divisée par la durée, aboutissant à deux mesures du niveau équivalent ($L_{EQ}$), en dB (SPL) et en dBA (pondéré). L'ensemble des 24 valeurs constitue le SMLT. Divers quantiles de la distribution d'énergie dans chaque canal (énergie mesurée toutes les 0,1 s) sont également fournis avec les valeurs spectrales moyennes.

Le SMLT est utilisé dans les domaines du bruit et de la voix, le plus souvent en échelle de fréquence linéaire (LÖFQVIST, 1986; NORDENBERG and SUNDBERG, 2004). S'il est établi sur une durée courte, de l'ordre de quelques syllabes, le SMLT s'avère extrêmement variable. Mais il se stabilise dans la durée et devient indépendant du contenu phonétique de la séquence orale enregistrée. Il reflète alors certaines caractéristiques de la voix du locuteur. La durée requise est normalement d'une quarantaine de secondes mais une durée plus courte (10 à 20 secondes) est suffisante pour comparer entre elles des séquences orales de même contenu phonétique.

Le travail de Pearsons et al. a été récemment reproduit (CUSHING et al., 2011) sur les mêmes bases (enregistrement étalonné en chambre sourde, 50 locuteurs anglophones, même dispositif d'analyse en 24 canaux). Une définition plus précise de la consigne vocale a été mise en œuvre, et une catégorie "hushed" (voix feutrée) a remplacé la catégorie "casual", produisant des niveaux d'environ 5 dB plus faibles. Ces données ne sont pas publiques. Les auteurs estiment que leurs mesures sont en moyenne très proches (à moins de 2 dB près) de celles de Pearsons et al, confirmant tout l'intérêt de celles-ci en tant que référence pour les études de l'effort vocal.

## 2.2 Variations individuelles du SMLT

La figure 1 montre deux exemples des SMLT produits par un locuteur (à gauche) et une locutrice (à droite) de même classe d'âge, selon les 5 consignes vocales proposées; par ordre de FDV croissante: "casual" (trait vert), "normal" (noir), "raised" (bleu), "loud" (violet) et "shout" (rouge). Le niveau global en dB SPL est indiqué avec la même couleur sur chaque figure. Chaque canal est désigné par son numéro; pour faciliter l'interprétation fréquentielle des tracés certains traits verticaux sont renforcés; les abscisses 4, 7 et 21 correspondent respectivement à 100, 200 et 5000 Hz, les abscisses 11 et 17 à 500 et 2000 Hz et l'abscisse 14 correspond à 1000 Hz. La zone des fondamentaux usuels se trouve entre les abscisses 3 (80 Hz) et 9 (320 Hz) et la zone du second

formant est centrée sur les abscisses 13 à 17 (800 à 2000 Hz). Il faut noter qu'un maximum du SMLT ne représente pas la fréquence d'un fondamental ou d'un formant, mais la fréquence moyenne autour de laquelle il évolue tout au long de la séquence. Cette grandeur peut être qualifiée de "fréquence dominante" (F0d, F2d etc).

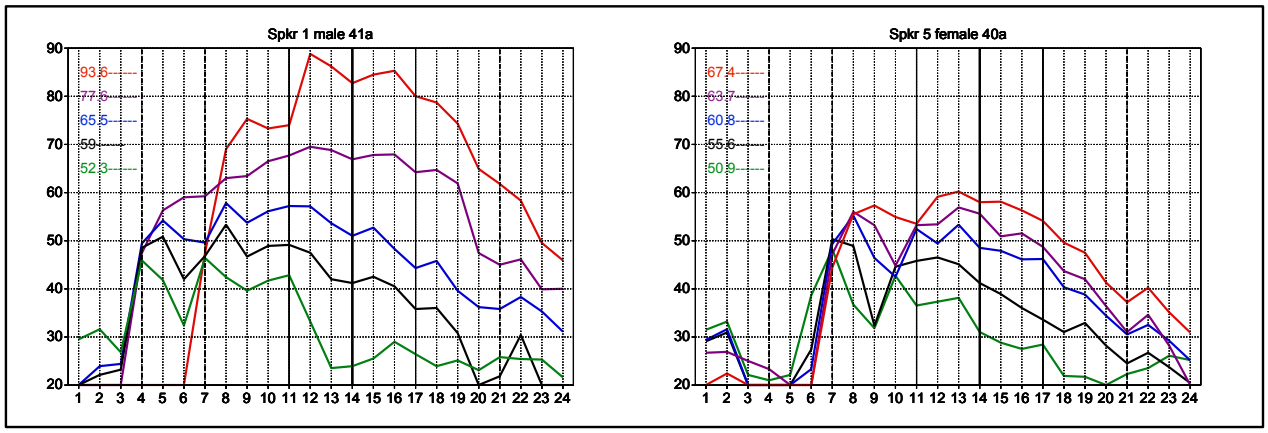



FIGURE 1: SMLT d'un locuteur (à g.) et d'une locutrice (à dr.) pour les 5 consignes vocales.

Ces deux exemples illustrent la grande variabilité des données. Pour une même consigne vocale deux locuteurs peuvent produire des FDV très différentes, variant de 25 dB et plus. Ainsi la production "shout" de la locutrice a une FDV de 67,4 dB et se trouve proche du mode "raised" du locuteur (à 65,5 dB) alors que le mode "shout" de ce dernier atteint 93,6 dB. Parallèlement à ces différences de FDV, les tracés des SMLT montrent une différence importante en basse fréquence, F0d se trouvant majoritairement au-dessous du canal 6 (160 Hz) pour le locuteur et au dessus pour la locutrice. Ceci à l'exception du mode "shout", pour lequel le locuteur atteint une F0d de l'ordre de 320 Hz, voisin de celui de la locutrice dans le même mode.

La progression d'un mode à l'autre s'accompagne de déformations d'ensemble: F0d se déplace vers l'aigu et le maximum spectral se déplace de F0d jusque vers 1000 Hz. La pente du spectre varie beaucoup selon la FDV. Pour les FDV moyennes l'amplitude décroît régulièrement de -6 à -8 dB/octave dans la partie haute du spectre. Le spectre des voix très fortes évolue de manière différente, la pente passant d'une valeur faible (0 à -3 dB/octave) dans la zone centrale (canaux 11 à 18) à une valeur forte (-12 à -18 dB/octave) dans l'aigu. Ces observations peuvent être mises en rapport avec les études théoriques des modèles d'onde glottique (DOVAL et al., 2006).

La grande différence entre les deux cas illustrés par la figure 1 n'est nullement exceptionnelle. En fait, les locuteurs ne respectent pas toujours la consigne vocale qui leur a été proposée. D'une manière générale les locutrices et les enfants ont tendance à produire des FDV maximales plus faibles que celles des locuteurs masculins. Il s'ensuit que la consigne vocale n'est pas un critère fiable de la FDV. Par contre, hors la zone de F0d, on observe une grande ressemblance des SMLT correspondant à une même FDV. Ceci permet de poser l'hypothèse selon laquelle la FDV, intensité émise, peut être estimée quantitativement à partir du seul profil du SMLT, indépendamment de l'intensité reçue.

## 2.3 A propos de la dynamique des enregistrements étalonnés en niveau

Les relevés indiquent que le rapport signal/bruit des enregistrements est proche de 100 dB ("dynamique d'enregistrement"). Cet intervalle doit être clairement distingué de l'intervalle séparant le niveau des voix les plus faibles de celui des voix les plus fortes ("dynamique de la voix"), qui excède rarement 50 dB. Dans la zone "conversationnelle" allant de "casual" à "loud"

l'intervalle est habituellement de 25 à 30 dB ("dynamique conversationnelle"). Il faut aussi prendre en considération le rapport signal/bruit propre à chaque SMLT ("dynamique propre" du SMLT), c'est-à-dire la différence de niveau entre le maximum spectral et le bruit de fond. La figure 1 montre un rebond d'intensité dans les canaux 1 et 2 (50 et 63 Hz), dans une zone où la voix n'a aucune énergie. Il s'agit de bruits parasites, qui limitent la dynamique propre des séquences les plus faibles et peuvent interférer avec le processus de normalisation et seuillage décrit plus loin.

## 2.4 Deux distributions selon le genre et l'âge des locuteurs

L'étude de Pearsons et al. permet de distinguer 3 catégories de locuteurs: hommes, femmes, enfants de moins de 13 ans. Les voix d'enfants et de femmes sont proches et il est légitime de les considérer ensemble, ce que nous ferons dans la suite. L'étude de certains indices comme les barycentres de diverses parties du SMLT montre l'existence de deux distributions distinctes. La figure 2 (à g), obtenue à partir des SMLT normalisés et seuillés entre 0 et 50 dB (cf section 3 ci-après), représente la fréquence du centre de gravité spectral en fonction de la FDV. Cette fréquence est exprimée en termes de numéros de canal (le canal 14 est à 1000 Hz).

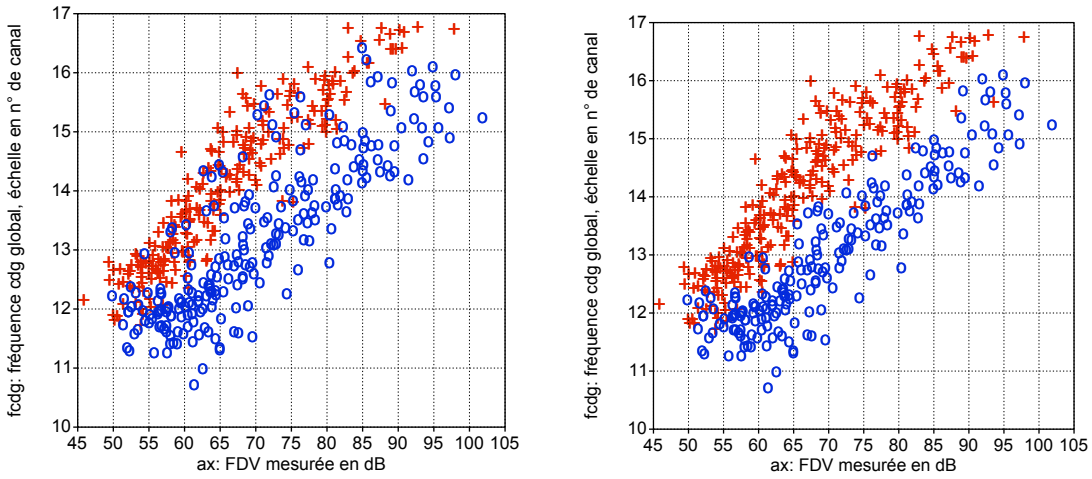



FIGURE 2: position du centre de gravité du SMLT en fonction de la force de voix.
En bleu: voix m adultes. En rouge: voix f et enfants <13 ans (à g), <16 ans (à dr).

On constate l'existence de deux nuages de points distincts (voix f et e en rouge, m en bleu) quasiment parallèles et croissants: à FDV constante le passage d'un groupe à l'autre entraîne un décalage du centre de gravité de 1,5 à 2 tiers d'octave. Un décalage comparable, de l'ordre d'une octave, s'observe à propos de la fréquence fondamentale dominante. Ces observations apparaissent massivement dans le SMLT à cause de l'échelle logarithmique qui dilate la zone basse du spectre au détriment de la zone haute.

Par ailleurs le rapport Pearsons définit les catégories "garçons" et "filles" par un âge maximum de 12 ans. La figure 2 (à g.) montre que de nombreux points provenant de voix masculines apparaissent dans le groupe des voix féminines. Il s'agit de jeunes locuteurs masculins. Du point de vue vocal il est donc plus réaliste de placer la limite à 15 ans, de façon à ce que le groupe des voix m ne comprenne que des hommes adultes, dont la mue est achevée. En procédant ainsi une trentaine de points initialement affectés au groupe masculin se trouvent réintégrés dans le groupe féminin (fig 2, dr.). Dans la suite nous considérerons donc seulement 2 groupes de locuteurs, le groupe m (hommes adultes >= 16 ans) et le groupe f+e (femmes adultes + enfants < 16 ans).

# 3 Estimation de la Force de Voix à partir du SMLT normalisé

Dans la section précédente il a été montré que la FDV est un meilleur descripteur du SMLT que ne l'est la consigne vocale donnée au sujet. Il s'agit maintenant de déterminer dans quelle mesure la FDV peut être prédite à partir du SMLT quand on retire de ce dernier l'information de niveau sonore absolu. La procédure adoptée consiste à normaliser tous les SMLT à un même niveau arbitraire, à comparer chacun à tous les autres, sauf ceux provenant du même locuteur, et à considérer la FDV du plus proche voisin comme l'estimation recherchée. La distance utilisée est de type L2 (moyenne quadratique des différences entre les valeurs spectrales des deux SMLT normalisés). Le résultat est exprimé sous deux formes: marge statistique d'erreur à 1 écart-type (dans une distribution normale 68% des observations se trouvent à moins de 1 écart-type de la moyenne), et coefficient de corrélation entre valeurs mesurées et valeurs estimées de la FDV.

## 3.1 Estimation de la FDV sur les données normalisées entre 0 et 50 dB

Tous les SMLT sont recalés en amplitude à un même niveau global de 50 dB. Les valeurs spectrales faibles sont limitées par un seuillage à 0 dB. Seuls les 20 canaux compris entre 3 (80 Hz) et 22 (6,3 kHz) sont pris en compte. La raison en est que l'utilisation de ces canaux extrêmes reviendrait à identifier les SMLT des voix faibles par leur faible recul du bruit de fond, qui dépend essentiellement des conditions de prise de son et non de la voix elle-même.

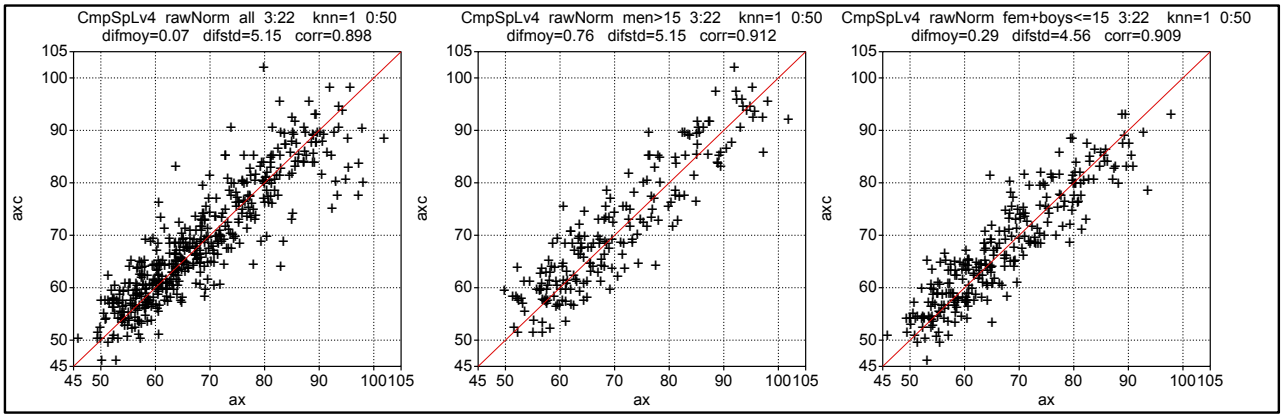



FIGURE 3: FDV estimée (axc), en fonction de la FDV mesurée (ax), pour le corpus total (fig 3a), pour les voix d'hommes adultes (3b) et pour les voix de femmes et d'enfants (3c).

Sur le corpus total (figure 3a) on observe une forte corrélation (0.898) et une marge statistique d'erreur de 5,15 dB. La distribution est mieux groupée dans la première moitié de l'échelle: les voix très fortes ou criées donnent des SMLT plus différents entre eux que les voix faibles ou moyennes. L'essentiel des erreurs grossières (à plus de 1 écart-type) concerne les voix très fortes ou criées, de FDV supérieure à 75 dB.

Les sous-corpus m (figure 3b) et f+e (figure 3c) donnent des résultats voisins, les marges d'erreur passant respectivement à 5,15 et 4,56 dB et les corrélations à 0,912 et 0,909. Les erreurs grossières sont moins nombreuses que dans le corpus total, ce qui suggère que la plupart d'entre elles sont imputables à la ressemblance des voix f+e moyennement fortes et des voix m très fortes ou criées.

Les résultats se dégradent peu quand on réduit la dynamique: la réduction à 30 dB au lieu de 50 dB augmente la marge d'erreur de 0,18 dB pour le corpus total, de 0,07 dB pour le sous-corpus m et de 0,48 dB pour le sous-corpus f+e. Il convient de rappeler que la dynamique est comptée à partir

du niveau global, lui-même supérieur d'une dizaine de dB au maximum spectral (ceci selon la forme du spectre). Entre le maximum spectral et le seuil on n'a guère plus d'une vingtaine de dB; ceci garantit que les performances obtenues reposent essentiellement sur les maxima spectraux et sont indépendantes du recul du bruit de fond des enregistrements.

### 3.2 Résultats avec plusieurs plus proches voisins

La quasi-symétrie des nuages de points autour de la diagonale est due au fait que l'on choisit comme valeur estimée de la FDV celle du SMLT le plus proche. L'ensemble des données n'étant pas infini on retrouve souvent en correspondance étroite la même paire de SMLT, l'un en test et l'autre en référence et réciproquement. La prise en compte de la distance moyenne des 4 plus proches voisins, pondérés selon leur classement, entraîne une amélioration globale (figure 4).

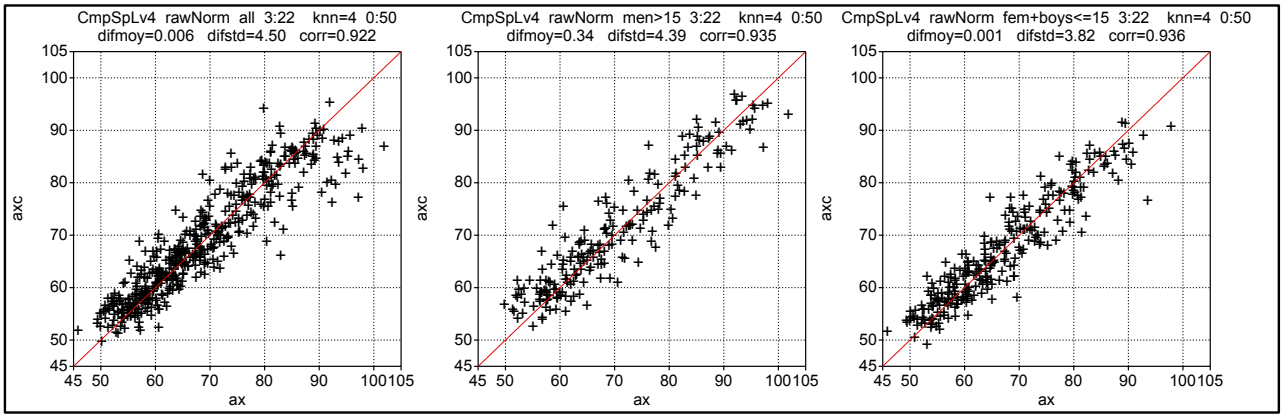



FIGURE 4: comme figure 3, mais la FDV est estimée à partir des 4 plus proches voisins.

Les marges d'erreur passent respectivement à 4,50, 4,39 et 3,82 dB et les coefficients de corrélation à 0,922, 0,935 et 0,936. Le nombre de valeurs surestimées (i.e. pour lesquelles la valeur estimée axc est très supérieure à la valeur mesurée ax) diminue et il subsiste pour le corpus total un certain nombre de valeurs sous-estimées. Globalement ces résultats valident l'hypothèse formulée plus haut, selon laquelle la FDV peut être prédite à partir de l'enveloppe du SMLT.

## 4 Discussion

Les données de Pearsons ne permettent pas de traiter complètement le problème d'estimation de l'EV. Il manque des nuances supplémentaires en voix faible, les mesures en mode "casual" ne portent pas sur le même texte que les autres, la prise de son à 1 m limite la dynamique des voix faibles, et surtout la non-disponibilité des enregistrements sonores empêche toute étude spectro-temporelle. Mais malgré leur ancienneté et leurs manques, ces données sont précieuses car elles constituent un matériau d'étude étalonné, indépendant du contenu phonétique, qui permet de relier directement la FDV à la forme du SMLT.

Les résultats sont encourageants, voire surprenants par leur relative précision, dans la mesure où l'information contenue dans le SMLT en 24 bandes - en fait 20 bandes dans les expériences rapportées ci-dessus - est très pauvre. Certaines erreurs sont imputables à la ressemblance spectrale entre la voix forte féminine et la voix masculine très forte ou criée, la F0d étant dans les deux cas de l'ordre de 300 Hz: s'il existe des différences significatives, celles-ci n'apparaissent pas suffisamment dans les SMLT; leur étude demanderait des données plus complètes. Il faut cependant souligner le fait que la voix criée pose des problèmes particuliers (ROSTOLLAND,

1982; JUNQUA, 1992; FUX et al., 2010), qu'elle est d'un emploi peu fréquent, et qu'il serait légitime de ne pas la considérer sur le même plan que la voix conversationnelle.

L'échelle logarithmique du SMLT en tiers d'octaves donne une grande importance aux fréquences inférieures à 1 kHz et sépare, au moins pour les FDV moyennes, le fondamental de l'harmonique 2. La différence d'amplitude entre ces deux composantes est depuis longtemps reconnue comme liée à l'EV, mais son utilisation en tant qu'indice est rendue problématique par la présence du premier formant F1 dans la même zone fréquentielle, de manière variable selon la voyelle, le genre du locuteur, et bien sûr la FDV. Le SMLT, qui intègre en 10 à 20 secondes une cinquantaine de syllabes, donne une représentation moyennée et acoustiquement stable de cette zone de fréquence, ce qui peut expliquer sa pertinence dans le problème d'estimation de la FDV.

Dans la suite il faudra confirmer les résultats et tenter de les étendre à d'autres ensembles de données, diverses langues, divers contenus phonétiques, diverses catégories de locuteurs. La limitation actuelle se trouve dans l'absence de bases de données étalonnées en termes de force de voix. La perspective, à terme, est de pouvoir affecter une valeur de la FDV à tout enregistrement de voix dont les conditions de prise de son (microphone, distance bouche-micro, gain) sont indéfinies, ce qui est le cas général.

L'estimation quantitative de la FDV concerne en particulier la recherche en acoustique phonétique. La variation de FDV produit d'importantes modifications spectro-temporelles du signal oral, qui perturbent la recherche d'invariants acoustiques associés aux éléments phonétiques et prosodiques de tous niveaux. Il faut prendre en compte la FDV, et donc pouvoir évaluer celle-ci. La même remarque vaut pour la prosodie, ainsi que pour le timbre individuel du locuteur ou les nuances liées à son expression.

Elle concerne également le traitement automatique de la parole et de la voix. Qu'il s'agisse de reconnaissance de la parole, du locuteur, ou du tour de parole, tous les systèmes se heurtent à la variabilité acoustique due à l'EV, qui fait chûter les performances. L'estimation préalable de la FDV permettrait d'en compenser les effets dans les processus d'apprentissage.

# 5 Conclusion

La notion qualitative d'effort vocal est une importante source de variabilité dans les sciences de la voix et de la parole. Des données de métrologie acoustique datant de 1977, réhabilitées récemment, permettent de démontrer l'intérêt du spectre moyen à long terme pour retrouver par le calcul, à moins de 5 décibels près, l'intensité acoustique émise par le locuteur, appelée Force de Voix. Dans le futur, la connaissance de cette grandeur devrait permettre d'expliquer, voire de compenser la variabilité qu'elle produit dans le signal oral. Pour progresser dans cette direction il est nécessaire de disposer de bases de données étalonnées en niveau sonore, qui malheureusement n'existent pas à ce jour.

# Remerciements

Un grand merci à Brian Katz (LAM, UPMC, Paris) qui nous a signalé l'existence de l'étude de Pearsons et al., à Albert Rilliard (LIMSI) pour de nombreuses conversations fructueuses, et à Anthony Nash (Charles Salter Associates, San Francisco) qui a réhabilité les données de Pearsons et les a mises à disposition de la communauté scientifique.

# Références


CUSHING I.R., LI F.F., COX T.J., WORALL K., JACKSON, T. (2011): "Vocal effort levels in anechoic conditions", Applied Acoustics, 72, 695-701.

DOVAL, B., D'ALESSANDRO, C. HENRICH, N. (2006). "The spectrum of glottal flow models", Acustica united with Acta Acustica, 92:1026-1046.

FUX, T., FENG, G., ZIMPFER, V. (2010). "Le rôle de la prosodie dans la perception de l'effort vocal", 10ème Congrès Français d'Acoustique, Lyon.

HANSON, H. (1997). "Glottal characteristics of female speakers: acoustic correlates", J. Acoust. Soc. Am. 101 (1), 466-481, 1997.

HUBER, J.E., STATHOPOULOS, E.T., CURIONE, G.M., ASH T.A. AND JOHNSON, K. (1999). "Formants of children, women, and men: the effects of vocal intensity variation", J. Acoust. Soc. Am. 106 (3), 1532-1542.

JUNQUA, J.-C. (1992). "The Lombard reflex and its role on human listeners and automatic speech recognizers", J. Acoust. Soc. Am. 93, 510-524.

LIENARD, J.S. AND DI BENEDETTO, M.G. (1999). "Effect of vocal effort on spectral properties of vowels", J. Acoust. Soc. Am. 106 (1), 411-422.

LIENARD J.S. AND BARRAS C. (2013). “Fine-grain voice strength estimation from vowel spectral cues”, InterSpeech, Lyon.

LIENARD J.S. (2014). “Etude des voyelles et de la force de voix par analyse discriminante”, JEP 2014, Le Mans.

LÖFKVIST A. (1986), "The long-time-average spectrum as a tool in voice research". Journal of Phonetics 14:472

NASH, A. (2014): "An electronic database of speech sound levels", Inter-Noise, Melbourne, 2014.

NORDENBERG M. AND SUNDBERG J. (2004), "Effect on LTAS of vocal loudness variation", Logoped Phoniatr Vocol 29, 183:191

PEARSONS K.S., BENNETT R.L., FIDELL S. (1977): "Speech levels in various noise environments", (Report No. EPA-600/1-77-025), U.S. Environmental Protection Agency, Washington DC.

RILLIARD, A., D'ALESSANDRO, C. AND EVRARD, M. (2018). " Paradigmatic variation of vowels in expressive speech: Acoustic description and dimensional analysis", J. Acoust. Soc. Am. 143, 109-122.

ROSTOLLAND, D. (1982). "Acoustic features of shouted voice", Acustica, vol 50, 118-125.

TRAUNMULLER, H. AND ERIKSSON, A. (2000). "Acoustic effects of variation in vocal effort by men, women and children", J. Acoust. Soc. Am. 107 (6), 3438-3451.

XXXIIe Journées d'Études sur la Parole
4-8 juin 2018, Aix-en-Provence, France

# Représentations de phrases dans un espace continu spécifiques à la tâche de détection d'erreurs.

Sahar Ghannay Nathalie Camelin Yannick Estève
LIUM, Le Mans Université, France
firstname.lastname@univ-lemans.fr

RÉSUMÉ

Cet article présente une étude sur la modélisation des erreurs de reconnaissance de la parole au niveau de la phrase, afin de compenser certains phénomènes mis en avant par l'analyse des sorties du système de détection d'erreurs que nous avons précédemment proposé. Nous avons étudié trois approches différentes, qui sont fondées respectivement sur l'utilisation des représentations continues (*embeddings*) de phrases dédiées à la tâche de détection d'erreur, d'un modèle contextuel probabiliste (MCP) et d'un réseau de neurones récurrent BLSTM. Une approche pour construire les *embeddings* spécifiques à la tâche est proposée et comparée à l'approche Doc2vec. Les expériences sont effectuées sur des transcriptions automatiques du corpus ETAPE générées par le système de reconnaissance automatique du LIUM. Elles montrent que les *embeddings* spécifiques à la tâche obtiennent de meilleurs résultats que les *embeddings* génériques et que leur intégration dans notre système améliore les résultats par rapport aux MCP et BLSTM.

ABSTRACT

**Task specific sentence embeddings for ASR error detection**

This paper presents a study on the modeling of automatic speech recognition errors at the sentence level. We aim in this study to compensate certain phenomena highlighted by the analysis of the outputs generated by the ASR error detection system we previously proposed. We investigated three different approaches, that are based respectively on the use of sentence embeddings dedicated to ASR error detection task, a probabilistic contextual model (PCM) and a bidirectional long short term memory (BLSTM) architecture. An approach to build task-specific sentence embeddings is proposed and compared to the Doc2vec approach. Experiments are performed on transcriptions generated by the LIUM ASR system applied to the ETAPE corpus. They show that the proposed sentence embeddings dedicated to ASR error detection achieve better results than generic sentence embeddings, and that the integration of task-specific embeddings in our system achieves better results than the PCM and BLSTM models.

MOTS-CLÉS : détection d'erreur, reconnaissance de la parole, réseau de neurones, représentation continue de phrase.

KEYWORDS: error detection, speech recognition, neural network, sentence embeddings.

## 1 Introduction

Les récentes avancées scientifiques dans le domaine du traitement automatique de la parole ainsi que la disponibilité de dispositifs de calcul puissants, ont conduit à l'obtention de performances


10.21437/JEP.2018-72

acceptables d'un point de vue applicatif dans le domaine de la reconnaissance automatique de la parole (RAP). Cependant, malgré ces performances, les sytèmes de RAP (SRAP) génèrent encore des erreurs de mots dans les transcriptions automatiques. Cela s'explique notamment par leurs sensibilités à la variabilité : d'environnement acoustique, de locuteur, de style de langage, de la thématique du discours, *etc*. Ces erreurs présentent un obstacle à l'application de certains traitements automatiques tels que l'extraction d'information, la traduction de la parole, la compréhension de la parole,*etc*.

Depuis deux décennies, de nombreuses études se focalisent sur la détection des erreurs de SRAP. Habituellement, les meilleurs systèmes de détection d'erreurs sont fondés sur l'utilisation des champs aléatoires conditionnels (CRF) comme dans (Parada *et al.*, 2010) et (Béchet & Favre, 2013). Des travaux récents ont commencé à appliquer les réseaux de neurones pour la tâche de détection d'erreurs. Dans ces travaux (Tam *et al.*, 2014; Ogawa & Hori, 2017), différentes architectures neuronales ont été exploitées : perceptrons multi-couches, réseaux de neurones récurrents, *etc.* Dans nos études précédentes (Ghannay *et al.*, 2015c, 2016a,b), nous avons étudié l'utilisation de différents types d'*embeddings* de mot. Dans (Ghannay *et al.*, 2015c), nous avons proposé une approche neuronale pour la détection d'erreurs dans les transcriptions automatiques et pour la calibration des mesures de confiance issues d'un SRAP. Nous avons également étudié la combinaison de différents types d'*embeddings* afin de tirer profit de leurs complémentarités. Le système de détection d'erreurs proposé inclut comme sources d'information : les *embeddings* de mots linguistiques, les descripteurs syntaxiques, lexicaux et prosodiques ainsi que des informations contextuelles extraites des mots voisins. Nous avons également enrichi notre système par des *embeddings* acoustiques de mots. L'utilisation de ces derniers en plus des autres descripteurs a amélioré les performances du système de détection d'erreurs proposé (Ghannay *et al.*, 2016a).

Dans cet article, nous présentons tout d'abord un résumé de nos études précédentes. Nous rappelons les performances obtenues par notre système de détection d'erreurs ainsi qu'une partie des résultats de l'étude sur l'analyse d'erreurs de ce système. Ensuite, pour compenser certains phénomènes mis en avant par cette analyse, nous proposons une étude sur la modélisation de l'erreur de reconnaissance au niveau de la phrase. Nous avons étudié trois approches différentes, fondées respectivement sur : l'utilisation des *embeddings* de phrases dédiées à la tâche de détection d'erreurs, un modèle contextuel probabiliste (MCP), et un réseau de neurones récurrent BLSTM (*bidirectional long short term memory*). Une approche pour construire les *embeddings* de phrases spécifiques à la tâche de détection d'erreurs est proposée et comparée à l'approche Doc2vec.

# 2 Système de détection d'erreurs

Le système de détection d'erreurs s'appuie sur une architecture neuronale fondée sur une stratégie multi-flux pour l'apprentissage d'un réseau de neurones, nommée Perceptron Multicouche Multi-Stream (*MLP-MS*).Une description détaillée de cette architecture est présentée dans (Ghannay *et al.*, 2015b).

## 2.1 Ensemble de descripteurs

Le système *MLP-MS* doit attribuer une étiquette *correct* ou *erreur* à chaque mot en l'analysant dans son contexte. Cette attribution est faite en s'appuyant sur l'ensemble de descripteurs suivants, dont certains sont identiques à ceux présentés dans (Béchet & Favre, 2013), pour chaque mot :

— Probabilités *a posteriori* générées par le SRAP.
— Descripteurs lexicaux extraits des sorties de SRAP : longueur du mot (nombre de lettres) et trois indices binaires indiquant si les trois 3-grammes contenant le mot courant ont été vus dans le corpus d'apprentissage du modèle de langue du SRAP.
— Descripteurs syntaxiques fournis par la boite à outils MACAON [1] appliquée aux sorties de SRAP. Des analyseurs morphosyntaxiques et de dépendances sont utilisés pour extraire les étiquettes syntaxiques, le gouverneur du mot courant et les liens de dépendance entre le mot courant et son gouverneur.
— Descripteurs prosodiques : le nombre de phonèmes, la durée moyenne des phonèmes, la durée de la pause précédant le mot sont extraits à partir de l'alignement forcé des transcriptions avec le signal audio. Ces paramètres sont détaillés dans (Ghannay *et al.*, 2015c).
— Le mot. Dans MLP-MS, il est représenté par son *embeddding* linguistique qui correspond à la combinaison par auto-encodeur de trois *embedddings* différents : *w2vf-deps* (Levy & Goldberg, 2014), *skip-gram* fourni par *word2vec* (Mikolov *et al.*, 2013), et *GloVe* (Pennington *et al.*, 2014). Cette combinaison est décrite dans (Ghannay *et al.*, 2015c). La représentation orthographique du mot est utilisée dans le système à base de CRF (Béchet & Favre, 2013).
— Les *embeddding* acoustiques. Ils correspondent aux *embedddings* acoustiques de signal et acoustiques de mot décrits dans (Ghannay *et al.*, 2016a).

## 2.2 Expériences et résultats

### 2.2.1 Données expérimentales

Les données expérimentales sont issues du corpus français ETAPE (Gravier *et al.*, 2012), composé d'enregistrements audio d'émissions télévisées (Broadcast News) et de leurs transcriptions manuelles. Ce corpus est enrichi avec des transcriptions automatiques générées par le système *LIUM SRAP*, qui est un système multi-passes basé sur le décodeur CMU Sphinx, utilisant des modèles acoustiques GMM/HMM. Ce système a gagné la campagne d'évaluation ETAPE en 2012. Une description détaillée est présentée dans (Deléglise *et al.*, 2009).

Les transcriptions automatiques ont été alignées avec les transcriptions de référence en utilisant l'outil *sclite* [2]. À partir de cet alignement, chaque mot dans le corpus a été étiqueté *correct* ou *erreur*. La description des données expérimentales est présentée dans le tableau 1.

| Nom | #mots ref | #mots hyp | WER |
|---|---|---|---|
| Train | 349K | 316K | 25,3 |
| Dev | 54K | 50K | 24,6 |
| Test | 58K | 53K | 21,9 |

TABLE 1 – Nombre de mots de référence (*#mots ref*), nombre de mots générés par le SRAP LIUM (*#mots hyp*) et taux d'erreur mot (*WER*) de chacun des sous-corpus issu d'ETAPE.

1. http ://macaon.lif.univ-mrs.fr
2. http ://www.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm

### 2.2.2 Résultats expérimentaux

#### Performance du système MLP-MS

Cette section présente les résultats expérimentaux de notre système de détection d'erreurs *MLP-MS* ainsi que l'analyse de ses sorties en fonction de l'empan moyen de l'erreurs, et les compare à un système état de l'art basé sur les CRFs implémentés avec *Wapiti* [3]. Les résultats sont évalués en termes de rappel (R), précision (P) et F-mesure (F) pour la détection de mots erronés et de taux d'erreur de classification globale (CER).

| **Corpus** | **Approche** | **Etiquette *erreur*** | | | **Global** |
|---|---|---|---|---|---|
| | | **P** | **R** | **F** | **CER** |
| Dev | CRF | 0,70 | 0,55 | 0,62 | 10,12 |
| | MLP-MS | 0,72 | 0,60 | 0,65 | **9,38** |
| Test | CRF | 0,69 | 0,54 | 0,61 | 8,46 |
| | MLP-MS | **0,70** | 0,61 | 0,65 | **7,75** |

TABLE 2 – Comparaison des performances des systèmes MLP-MS et CRF.

Les résultats expérimentaux présentés dans le tableau 2, montrent que notre système MLP-MS obtient de meilleurs résultats sur la détection de l'étiquette *erreur* et améliore significativement [4] les résultats par rapport à l'approche CRF.

#### Empan moyen d'erreurs

Dans cette section, nous nous intéressons à l'analyse des sorties de notre système de détection d'erreurs MLP-MS afin de percevoir les erreurs de reconnaissance qui sont difficiles à détecter. Ces analyses sont réalisées sur le corpus Dev en fonction de l'empan moyen de l'erreurs, *i.e.* une suite contiguë de mots erronés (Ghannay *et al.*, 2015a). Le tableau 3 présente la taille moyenne (en nombre de mots) des empans et l'écart type pour la vérité terrain, les prédictions et les prédictions correctes de deux systèmes : MLP-MS et CRF.

| **Corpus** | **Approche** | **Provenance du mot** | **Taille moyenne de l'empan** | **Écart type** |
|---|---|---|---|---|
| Train | | vérité terrain | 3,03 | 1,72 |
| Dev | | | 3,24 | 2,15 |
| Dev | CRF | prédictions | 3,28 | 1,77 |
| | | prédictions correctes | 2,88 | 1,34 |
| | MLP-MS | prédictions | 2,82 | 1,28 |
| | | prédictions correctes | 2,66 | 1,05 |

TABLE 3 – Empan moyen et écart-type pour la vérité terrain, les prédictions et les prédictions correctes de MLP-MS et CRF

3. http ://wapiti.limsi.fr
4. Un intervalle de confiance à 95% a été calculé et les résultats significativement meilleurs ont été soulignés.

On observe que la taille moyenne de l'empan des prédictions du CRF est proche de celle de la vérité terrain. En revanche, celle du système MLP-MS est, elle, plus petite de $12,9\%$ par rapport à la vérité terrain avec un écart type plus petit de $40,5\%$. Les prédictions correctes, tant pour celles produites par le système CRF que par celui proposé, présentent des empans de taille bien inférieure à la vérité terrain.

Nous supposons que l'écart lié à la taille de l'empan d'erreur entre la vérité terrain, les prédictions et les prédictions correctes, est dû à l'architecture de MLP-MS. Contrairement aux CRF, dont le décodeur produit des séquences d'étiquettes à partir d'un calcul sur l'ensemble de la séquence d'entrée, notre système prend ses décisions en s'appuyant sur un contexte local restreint.

S'appuyant sur cette observation, nous avons choisi d'explorer trois approches différentes détaillées dans ce qui suit.

# 3 Intégration d'informations globales à la phrase pour la détection d'erreurs

## 3.1 Représentation de la phrase dans un espace continu

Dans cette section, nous proposons d'enrichir notre système MLP-MS par des informations caractérisant la phrase, en exploitant des *embeddings* de phrases. Ces derniers ont été utilisés avec succès dans les tâches de classification de phrases et l'analyse de sentiments (Le & Mikolov, 2014; Tang *et al.*, 2016). Ces représentations peuvent être apprises d'une manière générale en utilisant l'outil *Doc2vec* (Le & Mikolov, 2014), ou d'une manière spécifique à la tâche, comme dans (Ren *et al.*, 2016).

### 3.1.1 Embeddings généralistes

Ce premier type *d'embeddings* de phrases s'appuie sur la méthode de sacs de mots distribués DBOW (*Distributed bag of words*) fournie par *Doc2vec* (Le & Mikolov, 2014). L'architecture DBOW consiste à prédire des mots choisis aléatoirement en fonction du paragraphe auquel ils appartiennent. Elle est apprise sur le corpus ETAPE pour construire un *embedding* de $100$ dimensions pour chaque transcription automatique, nommé $Em_{DBOW}$.

### 3.1.2 Embeddings spécifiques à la tâche

Les $Em_{DBOW}$ portent une information sur la sémantique contenue dans les transcriptions, mais ne portent probablement pas d'information sur les erreurs de transcriptions. C'est pourquoi nous proposons de construire des *embeddings* de phrases spécifiques à la tâche de détection d'erreurs. Pour ce faire, nous proposons d'utiliser les *embeddings* extraits d'un réseau de neurones convolutif (CNN) appris pour la classification des transcriptions automatiques en deux catégories de phrases : *peu erronées* (PE) ou *très erronées* (TE). Les *embeddings* extraits du *CNN* sont ainsi susceptibles de capter des informations sur les erreurs.

Le CNN est appris sur les transcriptions d'ETAPE annotées en *peu erronées* ou *très erronées*. En effet, nous avons considéré arbitrairement une phrase comme très erronée si 20% des mots qui la composent sont incorrects. Les phrases comprenant moins de 20% de mots incorrects sont alors considérées comme peu erronées (ensemble incluant les phrases totalement correctes). Le CNN prend en entrée le vecteur de descripteurs de la phrase et attribue en sortie une étiquette *PE* ou *TE* globale à la phrase. Le vecteur de descripteurs correspond à la concaténation des vecteurs de descripteurs des mots qui la composent, décrits dans la section 2.1. Le CNN est composé de deux couches de convolution et de sous-échantillonnage, suivies par deux couches de neurones qui sont totalement connectées sous la forme d'un MLP. La couche juste avant la couche de sortie *Softmax* est utilisée comme *embedding* de phrases de 100 dimensions, nommé $Em_{CNN}$. Le CNN obtient 13,5% de taux d'erreur de classification des transcriptions sur Test.

### 3.1.3 Résultats

Nous résumons dans la table 4 les performances des *embeddings* de phrases $Em_{DBOW}$ et $Em_{CNN}$. Elles sont comparées à celles obtenues par le système MLP-MS sans *embedding* de phrase (table 2).

| **Corpus** | **Représentation de phrase** | **Étiquette *erreur*** | | | **Global** |
|---|---|---|---|---|---|
| | | **P** | **R** | **F** | **CER** |
| Dev | $Em_{DBOW}$ | **0,73** | 0,58 | 0,65 | 9,36 |
| | $Em_{CNN}$ | 0,72 | 0,60 | 0,66 | **9,26** |
| Test | $Em_{DBOW}$ | **0,72** | 0,57 | 0,60 | 7,72 |
| | $Em_{CNN}$ | **0,72** | 0,58 | 0,64 | **7,69** |

TABLE 4 – Performances des *embeddings* de phrases $Em_{DBOW}$ et $Em_{CNN}$ sur Dev et Test

On remarque que les deux types d'*embeddings* de phrases apportent une légère amélioration par rapport aux résultats de MLP-MS. L'*embedding* $Em_{CNN}$ obtient de meilleurs résultats que l'*embedding* $Em_{DBOW}$ avec 1,27% et 0,77% de réduction de CER par rapport aux résultats de MLP-MS, respectivement sur Dev et Test. Nous pouvons émettre l'hypothèse que les *embeddings* extraits du CNN ont capté une information utile sur l'erreur.

Ce système sera nommé désormais MLP-MS$_{Em_{CNN}}$.

## 3.2 Modèle contextuel probabiliste pour une décision globale

Une autre approche pour compenser les lacunes remarquées lors de notre analyse d'erreurs consiste à utiliser un modèle contextuel probabiliste (MCP) qui porte des informations sur la distribution d'erreurs. Nous espérons ici corriger le problème de la taille de l'empan d'erreurs mal capturée par notre système de détection.

Cette approche est similaire à celle utilisée par (Dufour *et al.*, 2014) pour la détection automatique de segments de parole spontanée dans des émissions télévisées. Les auteurs ont proposé d'étendre un processus de classification locale à l'aide d'un modèle contextuel probabiliste d'étiquetage de séquences qui prend en compte l'étiquetage (parole préparée *vs.* parole spontanée) des segments voisins dans une fenêtre de taille 3. Grâce à cette extension, l'étiquetage, qui était issu d'une succession de décisions locales, devient un processus global.

Nous proposons d'appliquer cette idée à notre approche pour la détection d'erreur. Jusqu'à présent, l'étiquetage en *erreur vs. correct* des transcriptions automatiques par notre approche neuronale consistait en autant de classifications indépendantes que de mots à étiqueter. En tenant compte des classifications locales des mots voisins dans une fenêtre contextuelle de taille 5 identique à celle de l'entrée de notre système de détection d'erreurs, nous espérons lisser au niveau de la phrase le résultats de ces classifications. Pour cela, un modèle probabiliste d'ordre $n$ de distribution d'erreurs est utilisé : ce modèle estime la probabilité que le mot courant soit erroné en fonction de la justesse des 4 mots qui l'entourent.

### 3.2.1 Résultats

Nous avons utilisé la boite à outils *OpenFst*[5] pour créer le modèle sur les transcriptions automatiques du corpus ETAPE et les sorties de deux systèmes de détection : le système de base MLP-MS et le système MLP-MS$_{Em_{CNN}}$ qui intègre des connaissances sur la phrase.

Les systèmes résultants sont nommés avec l'extension -MCP. Les résultats obtenus par cette approche pour la détection d'erreurs sont résumés dans la table 5.

| **Corpus** | **Approche** | **Étiquette *erreur*** | | | **Global** |
|---|---|---|---|---|---|
| | | **P** | **R** | **F** | **CER** |
| Dev | MLP-MS-MCP | **0,73** | 0,58 | 0,65 | 9,31 |
| | MLP-MS$_{Em_{CNN}}$-MCP | 0,73 | 0,60 | 0,65 | **9,23** |
| Test | MLP-MS-MCP | 0,72 | 0,59 | 0,65 | **7,67** |
| | MLP-MS$_{Em_{CNN}}$-MCP | 0,73 | 0,57 | 0,64 | 7,69 |

TABLE 5 – Performances du modèle contextuel probabiliste pour la détection d'erreurs.

On observe que l'application du MCP aux sorties du système MLP-MS permet une légère diminution du CER tant sur Dev que sur Test. celui-ci est comparable à celui obtenu par le système MLP-MS$_{Em_{CNN}}$ qui intègre des informations globales sur la phrase.

L'application du MCP aux sorties de MLP-MS$_{Em_{CNN}}$ n'a amélioré que légèrement les résultats sur Dev. Cela peut s'expliquer par le fait que ce système intègre déjà des connaissances sur la phrase, et l'information apportée par l'approche globale est redondante.

## 3.3 Réseau de neurones BLSTM

Récemment, certaines architectures neuronales se sont révélées efficaces pour faire le traitement des séquences (Sutskever *et al.*, 2014). Il est donc intéressant de comparer l'approche neuronale utilisée jusqu'à présent dans nos expériences avec l'utilisation d'une architecture BLSTM. Cette comparaison nous permet d'évaluer l'impact des représentations continues de phrase dans une architecture classique par rapport à l'utilisation d'une architecture conçue pour apprendre des informations contextuelles distantes. Ce type d'architecture a notamment été utilisé avec succès pour la tâche de détection d'erreurs de transcription dans (Ogawa & Hori, 2017).

5. http ://www.openfst.org/twiki/bin/view/FST/WebHome

Dans nos expériences, le BLSTM est composé de deux couches de 512 unités chacune : 256 unités dans chaque direction. Il intègre les descripteurs décrits dans la section 2.1. Les résultats sont résumés dans la table 6.

| **Corpus** | **Système** | **Étiquette *Erreur*** | | | **Global** |
|---|---|---|---|---|---|
| | | **P** | **R** | **F** | **CER** |
| Dev | BLSTM | 0,70 | 0,64 | 0,67 | 9,28 |
| Test | BLSTM | 0,69 | 0,63 | 0,66 | 7,83 |

TABLE 6 – Comparaison de l'architecture MLP-MS proposée à l'architecture BLSTM.

Lorsque l'on compare les résultats obtenus par MLP-MS (table 2) et les résultats du BLSTM, on remarque que ceux-ci sont comparables. Le système MLP-MS$_{Em_{CNN}}$ montre des performances légèrement meilleures que celles du BLSTM. Cela nous permet ainsi de confirmer notre hypothèse sur l'utilité de l'intégration des informations globales sur la phrase dans notre système MLP-MS afin d'améliorer une prise de décision locale.

# 4 Conclusion

Dans cet article nous avons présenté une étude sur la modélisation des erreurs de reconnaissance de la parole au niveau de la phrase, afin de compenser certains phénomènes mis en avant par l'analyse des sorties du système de détection d'erreurs que nous avons précédemment proposé. Nous avons étudié trois approches différentes, qui sont fondées respectivement sur l'utilisation des *embeddings* de phrases dédiées à la tâche de détection d'erreurs, d'un modèle contextuel probabiliste et d'un réseau de neurones récurrent BLSTM. Nous avons également proposé une approche pour construire les *embeddings* spécifiques à la tâche et les comparer à l'approche Doc2vec. Les expériences sont effectuées sur des transcriptions automatiques du corpus ETAPE générées par le système de reconnaissance automatique du LIUM. Elles montrent que les *embeddings* de phrase spécifiques à la tâche obtiennent de meilleurs résultats que les *embeddings* génériques. De plus, leur intégration dans notre système améliore les résultats par rapport à l'application du modèle contextuel probabiliste sur les sorties du système MLP-MS et également par rapport à l'utilisation d'un BLSTM.

# Remerciements

Ce travail a été partiellement financé par la commission européenne à travers le projet EUMSSI, sous le numéro de contrat 611 057, dans le cadre de l'appel FP7-ICT-2013-10. Ce travail a également été partiellement financé par l'Agence nationale française de recherche (ANR) à travers le projet VERA, sous le numéro de contrat ANR-12-BS02-006-01.

# Références

BÉCHET F. & FAVRE B. (2013). Asr error segment localization for spoken recovery strategy. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, p. 6837–6841.

DELÉGLISE P., ESTÈVE Y., MEIGNIER S. & MERLIN T. (2009). Improvements to the LIUM French ASR system based on CMU Sphinx : what helps to significantly reduce the word error rate ? In *Interspeech*, Brighton, UK.

DUFOUR R., ESTÈVE Y. & DELÉGLISE P. (2014). Characterizing and detecting spontaneous speech : Application to speaker role recognition. *Speech Communication*, **56**, 1–18.

GHANNAY S., CAMELIN N. & ESTÈVE Y. (2015a). Which asr errors are hard to detect ? In *Workshop Errors by Humans and Machines in multimedia, multimodal and multilingual data processing (ERRARE 2015)*, Sinaia (Romania).

GHANNAY S., ESTÈVE Y. & CAMELIN N. (2015b). Word embeddings combination and neural networks for robustness in asr error detection. In *European Signal Processing Conference (EUSIPCO 2015)*, Nice (France).

GHANNAY S., ESTÈVE Y., CAMELIN N. & DELEGLISE P. (2016a). Acoustic word embeddings for asr error detection. In *Interspeech 2016*, San Francisco (CA, USA).

GHANNAY S., ESTEVE Y., CAMELIN N. & DELÉGLISE P. (2016b). Evaluation of acoustic word embeddings. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, p. 62–66.

GHANNAY S., ESTÈVE Y., CAMELIN N., DUTREY C., SANTIAGO F. & ADDA-DECKER M. (2015c). Combining continous word representation and prosodic features for asr error prediction. In *3rd International Conference on Statistical Language and Speech Processing (SLSP 2015)*, Budapest (Hungary).

GRAVIER G., ADDA G., PAULSSON N., CARRÉ M., GIRAUDEL A. & GALIBERT O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

LE Q. V. & MIKOLOV T. (2014). Distributed representations of sentences and documents. In *ICML*, volume 14, p. 1188–1196.

LEVY O. & GOLDBERG Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, p. 302–308.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR*.

OGAWA A. & HORI T. (2017). Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks. *Speech Communication*, **89**, 70–83.

PARADA C., DREDZE M., FILIMONOV D. & JELINEK F. (2010). Contextual Information Improves OOV Detection in Speech. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, volume 14, p. 1532–1543.

REN Y., WANG R. & JI D. (2016). A topic-enhanced word embedding for twitter sentiment classification. *Information Sciences*, **369**, 188–198.

SUTSKEVER I., VINYALS O. & LE Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, p. 3104–3112.

TAM Y.-C., LEI Y., ZHENG J. & WANG W. (2014). Asr error detection using recurrent neural network language model and complementary asr. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, p. 2312–2316 : IEEE.

TANG D., WEI F., QIN B., YANG N., LIU T. & ZHOU M. (2016). Sentiment embeddings with applications to sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, **28**, 496–509.

# Un protocole de recueil de productions orales chez l'enfant préscolaire : une étude préliminaire auprès d'enfants bilingues

Marie Philippart de Foy[1, 2], Véronique Delvaux[1, 2], Kathy Huet[1], Morgane Monnier[1]
Myriam Piccaluga[1] & Bernard Harmegnies[1]
(1) Institut de Recherche en Sciences et Technologies du Langage, Service de Métrologie et Sciences du Langage, Université de Mons, Belgique
(2) Fonds National de la Recherche Scientifique, Bruxelles, Belgique
Marie.Philippartdefoy@umons.ac.be ; Bernard.Harmegnies@umons.ac.be

## RESUME

Notre recherche vise à étudier le développement phonologique et phonétique, en production de parole, d'enfants bilingues préscolaires ayant différentes combinaisons linguistiques. Notre objectif est d'évaluer l'évolution des habiletés de production et des réalisations phonétiques des enfants dans une approche translinguistique et écologique. Dans cette optique, nous avons élaboré un protocole expérimental original et adapté. Les productions orales d'enfants bilingues sont recueillies longitudinalement via une tâche de dénomination en français. Le corpus cible des structures phonologiques spécifiques et implique un ordre de présentation particulier des items, par âge d'acquisition croissant et complexité phonologique progressive. Soumises à des procédures de codage spécifiques dans Praat, les productions récoltées font l'objet d'une transcription phonétique fine permettant d'isoler des phénomènes particuliers. Nous présentons les résultats préliminaires des premières analyses qualitatives pour trois enfants bilingues français-italien sur base de deux recueils de production de parole pour chaque participant.

## ABSTRACT

**A protocol for collecting speech production data from toddlers: a preliminary study with bilingual children.**

Our research aims at studying phonological and phonetic development, in speech production, in bilingual toddlers with different language pairs. Our objective is to assess the evolution of children's speech production skills and phonetic realizations with a cross-linguistic and ecological approach. In this perspective, we have developed an original and adapted experimental protocol. Speech productions from bilingual toddlers are longitudinally collected via a self-developed word-naming task in French. The corpus targets specific phonological structures and involves a particular presentation order of the items, organised by increasing age of acquisition and progressive phonological complexity. Collected productions are subjected to specific coding procedures in Praat and are phonetically transcribed in order to identify specific phenomena. We present the preliminary results of our first qualitative analyses for three bilingual French-Italian toddlers based on two data collections for each participant.

MOTS-CLES : acquisition, bilinguisme, production, phonologie, phonétique.
KEYWORDS: acquisition, bilingualism, speech production, phonology, phonetics.


10.21437/JEP.2018-73

# Introduction

Le développement phonologique et phonétique typique des enfants bilingues d'âge préscolaire est similaire mais pas identique à celui de leurs pairs monolingues et comporte certaines spécificités. Actuellement, la littérature est arrivée à un consensus d'après lequel les enfants bilingues développeraient dès le départ deux systèmes phonologiques distincts mais en interaction (Keshavarz, Ingram, 2002). Dès lors, cette interaction entre les deux systèmes pourrait causer de potentielles influences interlinguistiques, aussi bien au niveau segmental que suprasegmental, et plus précisément des phénomènes d'accélération (Lleó et *al.*, 2003), de décélération (Kehoe, 2002) et de transfert (Fabiano-Smith, Barlow, 2010) dans l'acquisition de certaines structures. En outre, l'occurrence et la directionalité de ces effets d'interaction dépendraient aussi bien du degré d'exposition à chaque langue que du degré de similarité entre les deux systèmes phonologiques. Dès lors, les enfants préscolaires bilingues auraient des trajectoires et stratégies développementales qui leur seraient propres. En effet, des différences entre bilingues et monolingues ont été observées au niveau : (1) du rythme et de l'ordre d'acquisition des segments et des inventaires phonétiques (Lleó et *al.*, 2003), (2) des caractéristiques phonologiques des formes de sortie (Kehoe, 2015) et (3) des processus phonologiques dits « simplificateurs » ou PPS (Lin, Johnson, 2010). Par ailleurs, il existe de nombreux types de bilinguismes. De fait, un très grand nombre de combinaisons linguistiques sont possibles et qui plus est, le bilinguisme est une expérience multidimensionnelle ; par conséquent, il existe une grande diversité de profils bilingues. Il faut donc s'attendre à une plus grande variabilité inter- et intra-individuelle chez les enfants bilingues et de ce fait, la distinction entre ce qui constituerait une variation normale dans les productions précoces et un potentiel trouble de la parole et/ou du langage peut s'avérer problématique (Armon-Lotem et *al.,* 2015). Actuellement, il n'existe pas encore de conclusion satisfaisante à propos de ce qui pourrait être qualifié de développement phonologique et phonétique bilingue typique *vs* atypique.

Par ailleurs, l'évaluation du développement phonologique et phonétique en production de parole pose des défis méthodologiques, d'autant plus chez les enfants bilingues ; par conséquent il est possible de relever certaines lacunes dans les études existantes. En effet, un certain nombre d'entre elles ont consisté en des études de cas et/ou ont impliqué un recueil de données unique. De plus, les études portant sur les bilingues incluent habituellement une seule combinaison linguistique. En outre, il s'agit généralement d'échantillons de parole continue collectés alors que les enfants sont enregistrés dans des situations de jeu non structurées. Par ailleurs, lorsqu'il s'agissait de recueil de mots isolés, certains auteurs ont utilisé une tâche de dénomination existante ; d'autres, tels que MacLeod et *al.* (2014), ont créé leur propre outil mais se sont basés sur un nombre restreint de critères pour la sélection des mots à faire produire à l'enfant. Enfin, les autres compétences linguistiques (lexicale ou morpho-syntaxique) n'ont pas toujours été prises en compte et les analyses acoustiques ont été peu fréquentes. Des travaux récents examinant le contrôle moteur de la parole chez l'enfant ont cependant impliqué des analyses acoustiques et articulatoires ainsi que de nouveaux outils, tels que l'échographie linguale (Barbier et *al.*, 2012). Néanmoins, les remarques ci-dessus soulignent la nécessité d'élaborer un paradigme observationnel approprié afin de récolter et d'analyser objectivement des sons de parole chez les enfants bilingues d'âge préscolaire.

Sur base de ces constats théoriques et méthodologiques, notre recherche vise à étudier le développement phonologique et phonétique, en production de parole, d'enfants bilingues préscolaires ayant différentes combinaisons linguistiques et, plus précisément, à évaluer l'évolution de leurs habiletés de production et de leurs réalisations phonétiques dans une approche translinguistique et écologique. Nous nous focalisons sur trois combinaisons linguistiques incluant toutes le français et une deuxième langue dont le degré de similarité avec le français varie : (1) français-italien, (2) français-arabe et (3) français-mandarin. Notre objectif est d'étudier l'impact spécifique que pourrait avoir chaque combinaison linguistique sur le développement phonologique et phonétique en français des enfants. De fait, ces trois combinaisons impliquent des contrastes spécifiques entre les deux systèmes phonologiques au(x) niveau(x) segmental et/ou suprasegmental

et, dès lors, sont susceptibles d'engendrer différents effets d'interaction pouvant résulter en des phénomènes intéressants à observer dans les productions en français.

# Méthodologie

Afin de répondre aux objectifs et de pallier les lacunes méthodologiques observées ci-dessus, nous avons mis au point un paradigme observationnel spécifique impliquant différents outils pour recueillir des données complémentaires auprès de nos participants de manière longitudinale et plus précisément, à intervalles réguliers de 4 mois. Nous collectons des données auto-rapportées via deux questionnaires complétés par les parents dans le but d'établir un profil linguistique précis pour chaque participant et de documenter leur développement lexical et morpho-syntaxique. Parallèlement, nous récoltons des productions de parole auprès des enfants et plus précisément, des mots isolés en français, langue commune à tous nos participants, via un outil original élaboré par nos soins.

## Participants

Notre étude implique 20 bilingues simultanés initialement âgés entre 21 et 35 mois et exposés à l'une des trois combinaisons linguistiques pré-citées. Plus concrètement, nous avons actuellement recruté 11 bilingues français-italien, dont 5 filles et 6 garçons, 7 bilingues français-arabe, dont 3 filles et 4 garçons, et 2 bilingues français-chinois, dont 1 fille et 1 garçon. Les participants évoluent principalement dans deux configurations familiales : (1) des familles impliquant un couple mixte où chaque parent s'adresse à l'enfant dans sa langue selon le principe « un personne-une langue » (Ronjat, 1913) et les deux langues sont donc parlées à la maison et (2), des familles où les deux parents parlent la même langue à leur enfant, ce dernier étant exposé au français à la crèche – minimum 4 jours par semaine et au plus tard à partir de 6 mois – et éventuellement via la fratrie. Plus précisément, six familles du groupe français-italien et trois familles du groupe français-arabe impliquent un couple mixte[1], quatre familles du groupe français-italien et quatre familles du groupe français-arabe impliquent deux parents parlant respectivement l'italien et l'arabe avec leur enfant. Une famille du groupe français-italien est mono-parentale et le parent s'adresse en italien à l'enfant. Enfin, les deux bilingues français-chinois proviennent de familles où la mère s'adresse en chinois à l'enfant et le père, en français.

## Recueil de données auto-rapportées

Les données auto-rapportées (ci-après AR) sont obtenues via deux outils complétés par les parents. Premièrement, ceux-ci répondent à un questionnaire construit à partir d'instruments existants, et plus particulièrement les questionnaires ALEQ (Paradis et *al*., 2010), ALDeQ (Paradis, 2011) et PABiQ (Tuller, 2015). Ce questionnaire parental permet : (1) de récolter un maximum d'information sur les spécificités de l'expérience bilingue pour chaque participant[2], (2) de calculer, sur base des réponses, différents scores pouvant générer deux indices : l'indice de non-risque ou INR (basé sur l'INR développé par Tuller, 2015) et l'indice de dominance linguistique ou IDL (basé sur l'IDL développé de Almeida et *al.*, 2016). L'INR prend en compte les facteurs de risque pour l'apparition d'un trouble du langage : il est obtenu sur base d'informations sur les premiers jalons développementaux[3] de l'enfant ainsi que sur l'existence d'une potentielle inquiétude parentale et de difficultés langagières au sein de la famille. Sa valeur est fixe et un indice inférieur à 17 peut être interprété comme indiquant un développement atypique. Pour l'IDL, un score d'exposition est tout d'abord calculé pour chaque langue et ensuite, l'indice est obtenu en soustrayant le score d'exposition de la

[1] Dans le groupe français-italien, c'est la mère qui parle l'italien dans trois des six familles et le père dans les trois autres. Dans le groupe français-arabe, c'est la mère qui parle l'arabe dans une des trois familles et le père dans les deux autres.

[2] C'est-à-dire les pratiques langagières à la maison et dans d'autres contextes ainsi que le degré d'exposition aux deux langues.

[3] Plus précisément, il s'agit de l'âge du premier mot et des premières combinaisons de mots.

deuxième langue à celui du français[4]. Une valeur d'IDL comprise entre -6 et +6 correspond à un bilinguisme équilibré, une valeur supérieure à 6 à une dominance en français et une valeur inférieure à -6 à une dominance dans l'autre langue. La valeur de l'IDL est susceptible d'évoluer parallèlement aux changements survenant dans l'environnement linguistique de l'enfant ; c'est pourquoi ces données sont actualisées lors de chaque recueil. Deuxièmement, les parents remplissent des rapports parentaux, et plus spécifiquement, des adaptations des *MacArthur-Bates Communicative Development Inventories* (Fenson et *al.,* 1993) dans les deux langues de l'enfant (Kern, Gayraud, 2010 ; Caselli, Casadio, 1995 ; Tardif et *al*., 2008). Ces rapports renseignent sur le développement lexical et morphosyntaxique et permettent de calculer des scores tels que le total des mots et la longueur moyenne de l'énoncé produits par l'enfant.

## Recueil de productions orales en français

Pour récolter les productions orales des enfants, nous avons élaboré une tâche de dénomination de mots, insérée au sein d'un jeu avec un livre imagier impliquant l'enfant et l'expérimentateur, afin de cibler des mots précis et des structures phonologiques spécifiques dans un contexte ludique et interactif. Pour la sélection des mots, nous avons choisi des critères psycholinguistiques, phonologiques et structurels listés par ordre d'importance : (1) l'âge d'acquisition (ci-après AoA), sur base des normes d'AoA objectif de Chalard et *al.* (2003) et des rapports parentaux de Kern et Gayraud (2010), et (2) l'imageabilité des mots, la présence dans le corpus total de (3) tous les phonèmes du français, (4) toutes les consonnes du français en position initiale/médiane/finale dans le mot, (5) de groupes consonantiques (ci-après GC) dans différentes positions dans le mot et (6) de différentes structures syllabiques et longueurs de mots. Le corpus final inclut 3 items d'entrainement, 48 items *test* et 2 items *leitmotiv* ([ma.ja] et [wi.wi]) représentant des personnages de dessin animé et impliquant des glides en position intervocalique. Ensuite, nous avons décidé de présenter les mots dans un ordre spécifique sur base de deux critères : l'AoA et le niveau de complexité phonologique. Plus précisément, les mots sont organisés par AoA croissant, des mots acquis le plus tôt à ceux acquis le plus tardivement, et pour chaque tranche d'AoA, par complexité phonologique croissante. La complexité phonologique a été évaluée à partir de critères précis et opérationnalisés afin de générer un classement de complexité des mots. Exposés en Table 1, ces critères de complexité se situent à différents niveaux phonologiques et possèdent différents degrés auxquels une valeur spécifique a été attribuée. La valeur la plus élevée a été assignée à un critère du niveau intersegmental, c'est-à-dire la présence d'un GC de 3 consonnes en coda. De fait, une telle séquence est relativement peu fréquente en français et très complexe à prononcer pour un enfant.

| Niveau de complexité | Critères | Valeur assignée |
|---|---|---|
| Suprasegmental *ICSS* | Structure et longueur des mots : mono-/bisyllabiques avec réduplication (0) *vs.* bisyllabiques avec structure variée (1) *vs.* trisyllabiques (2) | 0-1-2 |
| Segmental *ICS* | Absence (0) *vs.* présence de voyelles nasales (0,5) | 0-0,5-1-2 |
| | Absence (0) *vs.* présence des fricatives /ʃ/, /z/, /ʒ/ en position d'attaque (1) *vs.* coda (2) | |
| | Présence d'une initiale vocalique (1) | |
| Intersegmental *ICIS* | Absence (0) *vs.* présence de GC de 2 consonnes (1) en position de coda (2) | 0-1-2-3 |
| | Absence (0) *vs.* présence de GC de 3 consonnes (2) en position de coda (3) | |

TABLE 1 – Critères de complexité phonologique

[4] Le score d'exposition est calculé à partir d'informations sur l'âge du début et la durée d'exposition, la fréquence et la diversité des contextes d'exposition précoce, l'utilisation actuelle des langues à la maison et dans d'autres contextes, le nombre de mois passés à la crèche et l'éventuel début de scolarisation. Il faut préciser que nous avons adapté la manière de calculer ces deux indices au profil spécifique de nos participants, c'est-à-dire des bilingues simultanés initialement d'âge préscolaire.

Pour chaque mot, nous avons ensuite calculé un indice de complexité phonologique[5] sur base de l'équation ci-dessous où, pour le mot *i*, $IC_i$ correspond à l'indice de complexité globale, $ICSS_i$ à l'indice de complexité suprasegmentale, $ICS_i$ à l'indice de complexité segmentale, $ICIS_i$ à l'indice de complexité intersegmentale et $max_j$ à la plus haute valeur de l'indice parmi tous les mots *j* du corpus.

$$IC_i = \left( \frac{ICSS_i}{\max_j ICSS_i} + \frac{ICS_i}{\max_j ICS_i} + \frac{ICIS_i}{\max_j ICIS_i} \right) / 3$$

Les mots ont été organisés en 8 séries de 6 items, par AoA et complexité croissants, et les items *leitmotiv* ont été insérés entre chaque série. Les images proviennent principalement des bases de données de Moreno-Martinez et Montoro (2012) et Brodeur et *al.* (2012). Par ailleurs, notre protocole est adaptatif à différents niveaux ; de fait, l'expérimentateur peut choisir de/d' : (1) faire produire le corpus en entier ou partiellement, (2) intégrer le(s) parent(s) à la tâche afin de faciliter la production de l'enfant et (3), adapter l'administration de la tâche et les consignes, alternant entre de la dénomination (l'enfant répond à la question « Qu'est-ce que c'est ? ») et de la répétition (l'enfant répète la cible produite par l'adulte), en fonction de l'âge et des capacités attentionnelles de l'enfant. Les enfants sont enregistrés à leur domicile, dans une pièce calme, au moyen d'un enregistreur audio-portable *Zoom H5* et d'un micro chant *Sennheiser E912 BK*.

### Pré-traitement des données

Nous avons élaboré un système de codage dans le logiciel *Praat* consistant à créer un objet *Textgrid* afin d'annoter l'enregistrement sur plusieurs niveaux. Plus précisément, nous avons défini cinq couches d'annotation à l'intérieur desquelles le fichier son a été segmenté sur base de critères précis. Ces cinq niveaux d'annotation sont : (1) le tour de parole, c'est-à-dire le changement de locuteur au cours de l'échange conversationnel ; (2) le groupe de souffle du mot-cible, c'est-à-dire le groupe de mots énoncés en un seul souffle au cours duquel l'enfant tente de produire l'item attendu ; (3) la transcription phonétique de la production orale de l'enfant au moyen de l'Alphabet Phonétique International (API) ; (4) la technique d'élicitation employée par l'adulte pour faire produire le mot-cible à l'enfant et (5) les éventuels commentaires sur la production de parole.

## Résultats préliminaires

Pour cette section sur les résultats, nous nous focalisons sur le groupe de bilingues français-italien (ci-après FR-IT) qui comporte 11 enfants. Premièrement, nous présentons brièvement les indices calculés sur base du questionnaire parental pour tous les participants de ce groupe. Ensuite, pour les données en production de parole, nous exposons de manière plus détaillée les résultats préliminaires pour 3 des 11 enfants présentant des profils linguistiques contrastés, sur base de l'IDL.

### Données auto-rapportées

Comme mentionné précédemment, le questionnaire parental permet de générer deux indices, l'INR et l'IDL. Le tableau ci-dessous (Table 2) reprend les résultats obtenus par tous les participants du groupe bilingue FR-IT. Les données ont été collectées à deux reprises, au Temps 1 (T1) et au Temps 2 (T2), pour tous les participants excepté les deux derniers auprès desquels il n'y a eu actuellement qu'un seul recueil. Les trois participants apparaissant en gris sont ceux sur lesquels nous nous focaliserons pour les analyses des productions orales au point suivant. Sur base de l'IDL calculé, la Figure 1 représente le nombre d'enfants ayant une dominance linguistique (ci-après DL) en français, une DL en italien ou un bilinguisme équilibré et ce, sur les deux recueils de données (les sujets pour lesquels les données ont été récoltées à une seule reprise ne sont pas représentés dans ce graphique). D'après les données reprises dans le tableau, il apparaît que l'IDL évolue pour la plupart des

[5] Notons toutefois qu'il serait possible d'également tenir compte de la deuxième langue dans l'assignation des valeurs aux différents critères de complexité afin de développer un indice encore plus spécifiquement adapté aux enfants bilingues.

participants et le graphique nous montre qu'un participant passe effectivement d'un bilinguisme dominant en français à un bilinguisme équilibré du premier au deuxième recueil. Par ailleurs, la confrontation des résultats obtenus pour l'IDL avec une analyse qualitative du profil linguistique de chaque enfant indique que cet indice reflète adéquatement la réalité et semble donc être fiable.

| Participants | Age | | INR | IDL | |
|---|---|---|---|---|---|
| | *T1* | *T2* | | *T1* | *T2* |
| Sujet 1 | 25 | 29 | 26 | 10 | 3 |
| Sujet 2 | 25 | 29 | 26 | 5 | 5 |
| Sujet 3 | 35 | 39 | 29 | -18,5 | -15 |
| Sujet 4 | 33 | 37 | 24 | -18 | -16,5 |
| Sujet 5 | 31 | 35 | 24 | 1 | 3,5 |
| Sujet 6 | 21 | 25 | 28 | 15 | 15 |
| Sujet 7 | 21 | 25 | 29 | 17,5 | 17 |
| Sujet 8 | 24 | 28 | 29 | 21 | 19 |
| Sujet 9 | 34 | 38 | 23 | -11 | -10,5 |
| Sujet 10 | 35 | / | 29 | 10 | / |
| Sujet 11 | 23 | / | 27 | -18 | / |

TABLE 2 – Indices INR et IDL aux T1 et T2

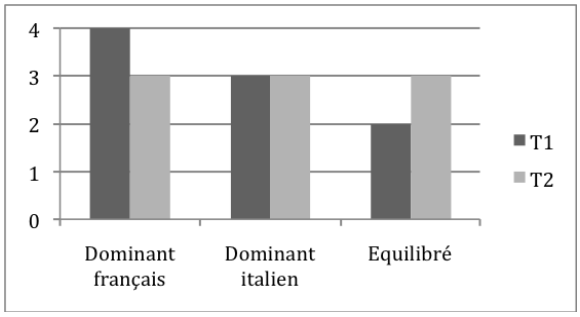



FIGURE 1 – Dominance linguistique des sujets aux T1 et T2

## Productions orales

Nous allons à présent comparer les données en production de parole de participants contrastés au niveau de leur DL afin d'observer l'évolution de leurs habilités de production d'un recueil à l'autre ainsi que d'examiner leurs points communs et spécificités. Nous avons donc sélectionné, dans l'échantillon FR-IT, trois participants dont les profils linguistiques diffèrent sur base de leur IDL. Plus spécifiquement il s'agit du sujet 2, bilingue équilibré, et des sujets 3 et 8 ayant respectivement une DL en italien et en français. Leurs caractéristiques sont résumées en Table 3 ci-dessous

| Participants | Age | | IDL | | Configuration familiale | Début de l'exposition aux langues | |
|---|---|---|---|---|---|---|---|
| | *T1* | *T2* | *T1* | *T2* | | | |
| Sujet 2 - ♂ | 25 | 29 | 5 | 5 | Couple mixte – mère italophone | naissance | |
| Sujet 3 - ♀ | 35 | 39 | -18,5 | -15 | Deux parents italophones | FR : 6 mois | IT : naissance |
| Sujet 8 - ♂ | 24 | 28 | 21 | 19 | Couple mixte – mère italophone | naissance | |

TABLE 3 – Tableau récapitulatif des caractéristiques des trois participants

Avant de comparer les phénomènes observés dans les productions orales des participants, nous allons au préalable préciser le nombre d'items produits par chacun, en distinguant les items dénommés des items répétés, pour les deux recueils. Ci-dessous, les Figure 2 et 3 représentent la proportion d'items dénommés *vs* répétés au T1 et T2.

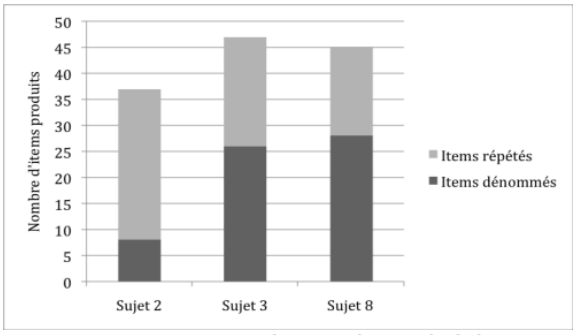



FIGURE 2 – Items dénommés *vs* répétés au T1

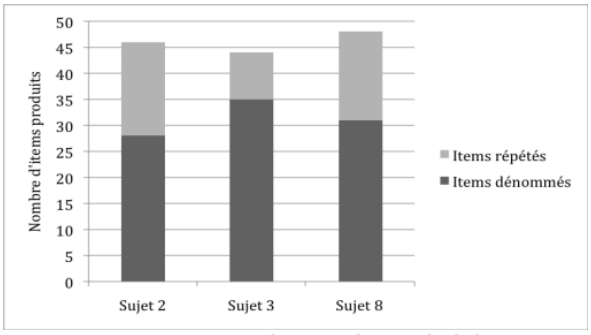



FIGURE 3 – Items dénommés *vs* répétés au T2

Pour les trois participants, le nombre d'items dénommés augmente et le nombre d'items répétés diminue d'un recueil à l'autre, et particulièrement chez le sujet 2. Il est important de le souligner car dans un processus de dénomination, l'enfant ne reproduit pas une cible énoncée par l'adulte alors que dans un processus de répétition, l'enfant a tendance à être dans une dynamique d'imitation verbale, surtout s'il s'agit d'un mot qui ne lui est pas familier. Dès lors, certaines caractéristiques des

productions orales de l'enfant peuvent être induites par l'adulte. Tous les items produits par les trois participants lors des deux recueils ont fait l'objet d'une transcription phonétique étroite au moyen des symboles de l'API sur base de laquelle une série de phénomènes intéressants ont pu être relevés pour les trois participants. Il s'agit principalement de changements affectant un ou plusieurs segment(s) que nous avons analysés en termes de processus phonologiques simplificateurs (ci-après PPS) classés en trois catégories: les PPS (1) structurels, (2) de substitution et (3) d'assimilation (d'après Maillart, 2006 ; Bishop, Minor-Corriveau, 2015). Pour chaque participant, les différents PPS ont été comptabilisés et sont repris dans le tableau ci-dessous (Table 4). Le tableau comporte également les occurrences de *code-switching* (ci-après CS), c'est-à-dire le recours à l'autre langue (l'italien), le nombre d'articles (indéfinis/définis) relevés dans les productions des enfants ainsi que le nombre total de mots et syllabes produits par chaque enfant lors des deux recueils.

| **Phénomènes relevés dans les productions orales** | **Sujet 2** | | **Sujet 3** | | **Sujet 8** | |
|---|---|---|---|---|---|---|
| *Processus phonologiques simplificateurs (PPS)* | *T1* | *T2* | *T1* | *T2* | *T1* | *T2* |
| **Processus structurels** | | | | | | |
| Suppression syllabe – Ex : parapluie [pa.ʁa.plɥi] => [pa.pɥi] | 14 | 8 | 5 | 3 | 8 | 1 |
| Suppression initiale vocalique – Ex : étoile [e.twal] => [twal] | 1 | 2 | 2 | 1 | 2 | 0 |
| Suppression consonne attaque – Ex : cuillère [kɥi.jɛʁ] => [ɥi.jɛʁ] | 0 | 4 | 4 | 3 | 6 | 0 |
| Suppression consonne coda – Ex : langue [lɑ̃g] => [lɑ̃] | 1 | 2 | 2 | 2 | 1 | 1 |
| Suppression-réduction GC – Ex : glace [glas] => [gas] | 11 | 12 | 10 | 7 | 11 | 12 |
| Ajout/insertion d'un segment – Ex : yaourt [ja.uʁt] => [ja. gutʁ] | 2 | 4 | 6 | 3 | 2 | 2 |
| Allongement vocalique – Ex : parc [paʁk] => [pa:k] | 1 | 4 | 4 | 3 | 3 | 3 |
| Réduplication syllabique – Ex : cadeau [ka.do] => [do.do] | 0 | 0 | 0 | 0 | 1 | 0 |
| Total processus structurels | 30 | 36 | 33 | 22 | 34 | 19 |
| **Processus de substitution** | | | | | | |
| Antériorisation de consonne – Ex : cheveux [ʃ ə.vø] => [sə.vø] | 4 | 5 | 2 | 5 | 6 | 10 |
| Postériorisation de consonne – Ex : cloche [klɔʃ] => [klɔç] | 1 | 4 | 10 | 8 | 6 | 5 |
| Occlusion – Ex : fleur [flœʁ] => [plœʁ] | 2 | 1 | 2 | 1 | 2 | 1 |
| Constriction – Ex : écharpe [e. ʃaʁp] => [e. ʃaʁf] | 0 | 1 | 0 | 0 | 2 | 2 |
| Voisement – Ex : vache [vaʃ] => [vaz] | 0 | 1 | 0 | 0 | 0 | 2 |
| Dévoisement – Ex : bateau [ba.to] => [pa.to] | 4 | 3 | 9 | 5 | 2 | 4 |
| Nasalisation des voyelles – Ex : téléphone [te.le.fɔn] => [tɛ̃.e.fɑ̃] | 2 | 0 | 3 | 0 | 1 | 0 |
| Oralisation des voyelles – Ex : pingouin [pɛ̃.gwɛ̃] => [pin. gwin] | 2 | 3 | 4 | 2 | 1 | 1 |
| Gliding – Ex : lit [li] => [ji] | 0 | 1 | 8 | 15 | 3 | 0 |
| Affrication– Ex : girafe [ʒi.ʁaf] => [dzi.ʁaf] | 0 | 0 | 1 | 0 | 0 | 2 |
| Déplacement (métathèse simple/double) – Ex : porte [pɔʁt] => [pɔtʁ] | 0 | 1 | 8 | 3 | 2 | 3 |
| Total processus de susbstitution | 15 | 20 | 47 | 39 | 24 | 30 |
| **Processus d'assimilation** | | | | | | |
| Harmonie consonantique – Ex : livre [livʁ] => [vivʁ] | 2 | 1 | 4 | 1 | 5 | 3 |
| Harmonie vocalique – Ex : pantalon [pɑ̃.ta.lɔ̃ ] => [pa.ta.lɔ̃] | 1 | 1 | 2 | 1 | 2 | 0 |
| Total processus d'assimilation | 3 | 2 | 6 | 2 | 7 | 3 |
| **Total de PPS relevés** | 48 | 58 | 86 | 63 | 65 | 52 |
| *Production de code-switching* | 0 | 3 | 1 | 6 | 0 | 0 |
| *Production d'un article défini/indéfini* | 0 | 3 | 35 | 29 | 11 | 29 |
| **Total de mots produits** | 37 | 46 | 47 | 44 | 45 | 48 |
| **Total de syllabes produites** | 57 | 85 | 124 | 108 | 85 | 121 |

TABLE 4 – Phénomènes relevés dans les productions orales des trois participants

Sur base du tableau, nous remarquons que la majorité des PPS sont communs aux trois participants, que le nombre total de PPS diminue pour les sujets 3 et 8 alors qu'il augmente pour le sujet 2 et que le sujet 3 présente le plus grand nombre de PPS. Si nous considérons chaque catégorie de PPS séparément, nous observons que : (1) les processus structurels diminuent chez les sujets 3 et 8 et augmentent chez le sujet 2, (2) les processus de substition diminuent chez le sujet 3 mais augmentent chez les sujets 2 et 8 et (3), les processus d'assimilation diminuent chez tous. Ensuite, on observe d'avantage de suppressions de syllabes chez les sujets 2 et 8 mais leur nombre diminue fortement d'un recueil à l'autre pour les deux enfants. Globalement et communément aux trois participants, le

PPS le plus fréquent est la réduction de GC et les PPS les moins fréquents sont les processus de réduplication, de voisement et d'affrication. Au cas par cas, il y a une occurrence élevée de suppression de syllabes chez le sujet 2, de gliding et de postériorisation chez le sujet 3 et d'antériorisation chez le sujet 8. En revanche, le sujet 2 ne présente pas, ou très peu, de phénomènes d'affrication et de déplacement, la constriction et le voisement n'apparaissent pas chez le sujet 3 et les phénomènes d'oralisation sont moins fréquents chez le sujet 8. Par ailleurs, il est important de noter que le nombre d'articles produits est le plus élevé chez le sujet 3 et qu'il augmente fortement entre le T1 et le T2 chez le sujet 8. Le sujet 2, quant à lui, en produit assez peu lors des deux recueils. Parallèlement, il y a d'avantage de phénomènes de CS chez le sujet 3, surtout au T2, que chez les sujets 2 et 8, ce dernier n'y ayant pas une seule fois recours.

# Discussion et conclusion

Nous avons présenté le protocole expérimental élaboré pour le recueil de productions orales d'enfants préscolaires dans le cadre de notre recherche visant à évaluer le développement phonologique et phonétique d'enfants bilingues ayant différentes combinaisons linguistiques. Nous avons ensuite exposé les résultats préliminaires pour trois bilingues FR-IT contrastés au niveau de leur DL. Avant de discuter ces résultats, il importe de constater que le protocole est efficace pour récolter les productions orales d'enfants, et qui plus est des productions de complexité phonologique variable. Les deux recueils ont également confirmé l'importance des données AR complémentaires permettant de caractériser précisément le bilinguisme de chaque participant ainsi que de considérer son développement langagier global. Sur base des transcriptions phonétiques, une série de phénomènes ont été observés et évalués quantitativement pour les trois participants ; il s'agit à présent de les analyser qualitativement en lien avec leur profil linguistique. Premièrement, le sujet 3 présente, lors des deux recueils, le plus grand nombre de PPS et l'occurrence la plus élevée de CS. De plus, il faut préciser que certaines de ses productions semblent dénoter un compromis entre les deux langues (par exemple : [dʒa.min] pour [pi.ʒa.ma]). Ces caractéristiques pourraient s'expliquer par sa DL en italien. Le sujet 8, quant à lui, présente très peu de phénomènes d'oralisation des voyelles nasales et aucun cas de CS n'est observé dans ses productions. Ces caractéristiques pourraient vraisemblablement découler de sa DL en français. En outre, l'augmentation importante du nombre d'articles produits par l'enfant indique une certaine évolution au niveau morphosyntaxique. Enfin, le sujet 2 parait être à un stade développemental moins avancé. Au T1, il produit environ dix items de moins que les deux autres sujets et ses productions se caractérisent par un grand nombre de suppressions de syllabes ; dès lors, le nombre total de syllabes qu'il produit est nettement moins élevé. En outre, son inventaire phonémique est plus restreint. Si le nombre de suppression de syllabe diminue de moitié au T2, les autres PPS ont tendance à augmenter. Ceci pourrait résulter, entre autres, de la complexification syllabique des productions de l'enfant. Enfin, nous avons observé, chez ce même sujet, trois occurrences de CS au T2, pouvant témoigner d'une certaine interaction entre les deux langues. Cependant, le CS peut être interprété de diverses manières et relève également du domaine sociopragmatique. Par ailleurs, nous avons pu observer, pour tous les participants, différentes phases développementales : jusqu'à 25 mois environ, les enfants sont majoritairement dans une dynamique d'imitation verbale et répètent l'item après l'adulte ; ce n'est que plus tard, après 30 mois, qu'ils commencent à dénommer spontanément et cette évolution pourrait se manifester, transitoirement, par une augmentation des PPS. Toutefois, il faudrait encore distinguer les PPS appliqués sur les items dénommés *vs* répétés.

Pour conclure, nous pouvons ajouter que cette étude préliminaire a également permis de poser les jalons d'une réflexion sur la méthode appropriée pour effectuer les transcriptions phonétiques et sur la nécessité d'avoir recours à plusieurs outils pour l'analyse des productions orales. L'étape suivante sera donc d'approfondir nos analyses phonologiques et phonétiques, entre autres en effectuant l'inventaire des structures phonologiques produites *vs* maîtrisés par chaque enfant ou en calculant des indices phonologiques développementaux, et bien évidemment, de procéder à des analyses acoustiques spécifiques et adaptées à la parole de l'enfant.

# Références


ARMON-LOTEM S., DE JONG J., MEIR N. (2015). *Assessing multilingual children: Disentangling bilingualism from language impairment*. Bristol: Multilingual matters.

BARBIER G., PERRIER P., MENARD L., BOË L. J. (2012). Contrôle lingual en production de parole chez l'enfant de 4 ans: une méthodologie associant étude articulatoire et modélisation biomécanique. In *14ème édition des Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, 1, 393-400.

BISHOP K., MINOR-CORRIVEAU M. (2015). Les processus phonologiques impliquant les groupes consonantiques en position initiale et finale: une étude sur l'articulation et la phonologie chez des enfants francophones et bilingues du nord de l'Ontario. In *Actes de l'ACFAS*, 21–61.

BRODEUR M. B., DIONNE-DOSTIE E., MONTREUIL T., LEPAGE M. (2010). The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PloS one*, 5(5), e10773.

CASELLI M. C., CASADIO P. (1995). *Il primo vocabolario del bambino: guida all'uso del questionario MacArthur per la valutazione della comunicazione e del linguaggio nei primi anni di vita*. Milan : FrancoAngeli.

CHALARD M., BONIN P., MEOT A., BOYER B., & FAYOL M. (2003). Objective age-of-acquisition (AoA) norms for a set of 230 object names in French: Relationships with psycholinguistic variables, the English data from Morrison et al.(1997), and naming latencies. *European Journal of Cognitive Psychology*, 15(2), 209-245.

de ALMEIDA L., FERRE S., MORIN E., PREVOST P., dos SANTOS C., TULLER L., ZEBIB R. (2016). L'identification d'enfants bilingues avec Trouble Spécifique du Langage en France. In *SHS Web of Conferences*, 27, 10005.

FABIANO-SMITH L., BARLOW J. A. (2010). Interaction in bilingual phonological acquisition: Evidence from phonetic inventories. *International journal of bilingual education and bilingualism*, 13(1), 81-97.

FENSON L., DALE P. S., REZNICK J. S., THAL D., BATES E., HARTUNG J. P., PETHICK S., REILLY J. S. (1993). *MacArthur Communicative Development Inventories: User's guide and technical manual*. San Diego : CA Singular Publishing Group.

KEHOE M. (2015). Lexical-phonological interactions in bilingual children. *First Language*, 35(2), 93-125.

KERN S., & GAYRAUD F. (2010). *L'inventaire français du développement communicatif*. Grenoble: La Cigale.

KESHAVARZ M., INGRAM D. (2002). The early phonological development of a Farsi-English bilingual child. *International Journal of Bilingualism*, 6(3), 255-269.

LLEO C., KUCHENBRANDT I., KEHOE M., TRUJILLO C. (2003). Syllable final consonants in Spanish and German monolingual and bilingual acquisition. in N. MULLER (Eds.), *(In)vulnerable Domains in Multilingualism, 191-220. Amsterdam*, Philadelphia: John Benjamins.

LIN L. C., JOHNSON C. J. (2010). Phonological patterns in Mandarin–English bilingual children. *Clinical linguistics & phonetics*, 24(4-5), 369-386.

MACLEOD A. A., SUTTON A., SYLVESTRE A., Thordardottir E., & TRUDEAU N. (2014). Outil de dépistage des troubles du développement des sons de la parole: bases théoriques et données préliminaires. *Canadian Journal of Speech-Language Pathology & Audiology*, 38(1), 40.

MAILLART C. (2006). Le bilan articulatoire et phonologique. L'évaluation du langage et de la voix, 26-51.

MORENO-MARTINEZ F. J, MONTORO P. R. (2012). An ecological alternative to Snodgrass & Vanderwart: 360 high quality colour images with norms for seven psycholinguistic variables. *PloS one*, 7(5), 37527.

PARADIS J., EMMERZAEL K., SORENSON DUNCAN T. (2010) Assessment of English Language Learners: Using Parent Report on First Language Development. *Journal of Communication Disorders*, 43,474-497.

PARADIS J. (2011) Individual Differences in Child English Second Language Acquisition: Comparing Child-Internal and Child-External Factors. *Linguistic Approaches to Bilingualism*, 1(3), 213-237.

RONJAT J. (1913). *Le développement du langage observé chez un enfant bilingue*, Paris: Champion.

TARDIF T., FLETCHER P., ZHANG Z. X., LIANG W. L. (2008). *Chinese communicative development inventories: User's guide and manual*. Peking University Medical Press.

TULLER L. (2015). Clinical use of parental questionnaires in multilingual contexts. In ARMON-LOTEM S., DE JONG J., MEIR N., *Assessing multilingual children: Disentangling bilingualism from language impairment*, Bristol: Multilingual Matters, 299-328.

# euh, rire et bruits en parole spontanée : application à l'alignement forcé

Brigitte Bigi, Christine Meunier
Laboratoire Parole et Langage, CNRS, Aix-Marseille Université
13100 Aix-en-Provence, France
brigitte.bigi@lpl-aix.fr, christine.meunier@lpl-aix.fr

## RESUME

Contrairement à la parole contrôlée, dans laquelle les intentions du locuteur sont très restreintes, la parole spontanée fait référence à une activité plus libre mais aussi plus riche de facteurs caractéristiques de l'interaction langagière. A ce titre, de nombreux phénomènes apparaissent comme les hésitations, les mots tronqués, les réductions phonétiques, etc. Nous proposons dans un premier temps un recensement de 3 événements paralinguistiques ("euh", rire, bruit), dans différents corpus spontanés : débat politique, narration, dialogue orienté tâche, dialogue informel avec consigne et dialogue informel sans consigne. Bien que ces événements soient fréquemment produits par les locuteurs, nous observons des différences significatives selon les corpus. A titre applicatif, nous montrons que les résultats de l'alignement forcé peuvent être nettement améliorés lorsque le système dispose d'un modèle acoustique qui inclut ces événements.

## ABSTRACT

**Filled pause, laughter and noise in spontaneous speech: application to forced-alignment.** Contrariwise to controlled speech, for which speaker's intention are very limited, spontaneous speech refers to a freer but also richer activity that is characteristic of language interaction. Many phenomena appear like hesitations, truncated words, phonetic reductions, etc. In this paper, we first propose a frequency survey of 3 paralinguistic events ("uh", laughter, noise), in different spontaneous corpora: political debate, interviews, task-oriented dialog, informal dialog with instructions and informal dialog without instructions. Even if these events are frequently produced by the speakers, we observe significant differences according to the corpora. For illustrative purposes, we finally show that the results of forced-alignment are significantly improved when the acoustic model of the system includes these events.

MOTS-CLES : parole spontanée, euh, rire, bruit, alignement forcé.
KEYWORDS: spontaneous speech, filled pause, laughter, noise, forced-alignment.

# 1 Introduction

La tâche d'alignement forcé consiste à déterminer automatiquement la localisation temporelle des phonèmes d'un fichier audio. Tandis que l'alignement de la parole lue a été largement abordée, les études relatives à la segmentation de la parole spontanée restent peu nombreuses. Les productions de parole extraites de situations naturelles et occasionnelles sont caractérisées par un débit de parole rapide mais aussi irrégulier, des troncatures de mots, des réductions de phonèmes (Johnson, 2004), etc. Le discours spontané est en effet produit dans une situation communicative dynamique impliquant des routines linguistiques et des contraintes qui conduisent à une réorganisation de la


10.21437/JEP.2018-74

production sonore, puis à des variations massives. Ces caractéristiques entraînent de grandes difficultés lorsque le flux vocal doit être annoté en unités phonétiques discrètes.

Les différents types de paroles spontanées peuvent fournir divers degrés de réduction, selon que la production est plus ou moins contrôlée. Une difficulté importante pour les outils d'alignement automatique réside dans le fait que la réduction n'est pas systématiquement une suppression discrète de phonèmes. Plusieurs études (Adda-Decker et al., 2013, Meunier, 2013) ont montré que, très souvent, la réduction entraîne une coalescence des phonèmes (plusieurs phonèmes sont fusionnés en un segment). Ces instances sont assez fréquentes et ne sont généralement pas perçues par les transcripteurs. De plus, la parole spontanée est caractérisée par plusieurs éléments qui n'apparaissent pas en condition contrôlée. En particulier rires, toux, bruits de bouche, etc. apparaissent fréquemment en conversation. Certains travaux (Ogden, 2001) indiquent que les *clics*, par exemple, sont utilisés de manière linguistique pour structurer le discours oral. Ces éléments n'appartiennent pas aux inventaires phonologiques, cependant, ils sont très présents dans les conversations par exemple et les outils automatiques doivent les identifier afin de fournir un alignement phonétique correct.

L'étude que nous proposons dans cet article s'inscrit dans le cadre de la linguistique de corpus. A partir des exemples contenus dans des données réelles, nous recensons différents événements de la parole spontanée. Nous avons ainsi sélectionné un ensemble de 5 corpus, relativement homogènes dans leur forme, de sorte que nous puissions observer et comparer les variations de fréquences des tokens. Nous nous focalisons sur 3 événements bien particuliers : le "euh", le rire et le "bruit". Dans un deuxième temps, nous montrerons où se situent ces événements relativement au reste de la parole : isolés entre deux silences, en début de segment, en fin de segment ou au sein d'un segment de parole. Nous avons ensuite évalué l'impact de la prise en compte de ces derniers dans la tâche d'alignement automatique. Pour ce faire, nous avons comparé les résultats d'un même système selon qu'il utilise un modèle acoustique qui inclut soit une représentation prototypique, soit un modèle spécifiquement appris pour chacun des trois événements euh, rire et bruit.

# 2 Corpus collectés

## 2.1 Origine et description des corpus

Pour cette étude, nous avons réuni et sélectionné les corpus décrits en table 1, afin qu'ils soient les plus homogènes possible. La première colonne indique le nom couramment donné au corpus, la deuxième le type d'enregistrement, la troisième la durée de parole (c'est-à-dire qu'elle n'inclut pas les silences), enfin, la quatrième colonne rapporte le nombre de locuteurs ainsi que le style de parole qui est détaillé entre parenthèses.

Dans la mesure où notre étude concerne spécifiquement certains événements paralinguistiques, il était important qu'aucun autre paramètre ne puisse influencer voire biaiser les résultats. Tous les corpus qui ont été sélectionnés ont été enregistrés en chambre sourde, avec un micro par locuteur. Chaque signal audio a été automatiquement segmenté en Unités Inter-Pausales (IPUs), i.e. des segments de production sonore alignés sur le signal. Ces segments sont entourés de silences dont la durée dépasse 200ms. Les frontières des IPUs ont été vérifiées manuellement pour chacun des corpus. Une transcription orthographique enrichie a ensuite été réalisée au sein des IPUs, en suivant

la convention de transcription spécifique au logiciel SPPAS[1] (Bigi, 2015). Ces transcriptions comprennent : les pauses pleines (euh), les rires (@), les bruits (*), les pauses courtes (+), les disfluences (répétitions, mots tronqués, …), les prononciations inhabituelles ainsi que les élisions inhabituelles. Compte-tenu des conditions d'enregistrement, dans les corpus de la table 1, les bruits se limitent à des productions du locuteur, à savoir des souffles, respirations, toux, etc. Enfin, dans *Europe* et *CID*, tous les euh, rires et bruits ont été alignés manuellement sur le signal.

| **Nom** | **Enreg.** | **Durée de parole** | **Loc.** | **Style de parole** |
|---|---|---|---|---|
| *Europe* | audio | 33 min | 6 | Débat politique à la radio |
| *Typaloc* | audio | 39 min | 4 | Conversation (interview) |
| *AixMapTask* | audio-vidéo | 163 min | 10 | Conversation (orientée tâche) |
| *CID* | audio-vidéo | 7h30min | 16 | Conversation (dialogue informel) |
| *Cheese* | audio-vidéo | 63 min | 8 | Conversation (dialogue informel) |

TABLE 1 : Description des corpus

Le corpus *Europe* (Portes, 2004) est un débat politique enregistré sur une station de radio essentiellement dédiée à l'information. Ce débat implique quatre invités interrogés par deux journalistes à propos de l'Union Européenne et plus particulièrement de la question délicate de ses frontières. Le corpus *TYPALOC* (Meunier *et al.*, 2016) original se compose de plusieurs corpus de lectures (mots et textes) et de parole spontanée (interviews), produites par des locuteurs en bonne santé et des locuteurs affectés par une dysarthrie. Pour cette étude, nous avons conservé uniquement les interviews (8-17 min) de locuteurs en bonne santé. La condition audio-visuelle du corpus *AixMapTask*[2] se compose d'enregistrements audio et vidéo de dialogues orientés tâche (Gorish *et al.*, 2014). Le protocole expérimental suit les règles standards d'une Map Task : les participants sont autorisés à dire tout ce qu'ils veulent dans le but d'accomplir la tâche qui leur incombe. Dans ce corpus, les participants sont assis face-à-face avec leurs cartes respectives posées sur des pupitres. Le *Corpus of Interactional Data - CID*[3] (Bertrand *et al.*, 2008) est constitué de dialogues d'une heure chacun, enregistrés en audio et en vidéo. L'un des deux sujets de conversation a été suggéré aux participants : des conflits dans leur environnement professionnel (5 dialogues) ou des situations insolites auxquelles ils ont été confrontés (3 dialogues). *Cheese* (Priego-Valverde & Bigi, 2016) est également un enregistrement audio-vidéo de dialogues (~ 15 min) impliquant deux participants. Il a été demandé à chacun de lire une blague, imposée par l'expérimentateur, puis de discuter librement. La partie relative à la lecture a été retirée du corpus pour la présente étude.

[1] http://www.sppas.org/. La version 1.9.4 a été utilisée pour cette étude.

[2] Enregistrements audio-vidéo et transcription orthographique sont disponibles sur Ortolang : https://hdl.handle.net/11403/sldr000875

[3] Enregistrements audio, transcription orthographique, et d'autres annotations sont disponibles sur Ortolang : https://hdl.handle.net/11403/sldr000720. Vidéo : https://hdl.handle.net/11403/sldr000027

## 2.2 Distributions des tokens

Chacun de ces corpus a été normalisé avec l'outil "Text Normalization" de SPPAS. La table 2 rapporte le nombre de "tokens" de chacun des corpus après cette normalisation (colonne 2). Par token, nous entendons toute séquence transcrite, à savoir les mots ainsi que toutes les autres productions sonores. Les colonnes 3 à 5 indiquent le pourcentage que représentent respectivement les "euh", rires et bruits. On constate que chacun des corpus contient un pourcentage relativement élevé de "euh" : de 2,3% à 6% des tokens. Pour les corpus *Europe* (6%) et *CID* (4%), cela fait du "euh" le token le plus fréquent, largement devant "de" avec ses 4,22% dans *Europe* et le mot "est" avec 2,67% dans le *CID*. Dans le corpus *Typaloc*, comme dans *Europe*, c'est le mot "de" qui est le plus fréquent (3,03%), tandis que dans *Cheese*, c'est le mot "est" comme pour le *CID*, avec 3,06%. Avec 5,36% des occurrences, "tu" est le mot le plus fréquent du corpus *AixMapTask*. On en conclut que « euh » est très fréquent en parole spontanée, cependant sa fréquence dépend du style de parole : on l'observe moins dans les différents styles de conversations que dans un débat politique radio-diffusé.

Concernant les rires, le corpus *Europe* n'en contient qu'un seul, ce qui n'est pas surprenant compte-tenu du type de débat et du sujet abordé. Le rire est en revanche relativement fréquent dans le *CID*. Toutefois, c'est dans *Cheese* qu'on retrouve proportionnellement le plus grand nombre de rires ; il arrive en 3ème position des tokens les plus fréquents. Ainsi, c'est dans les deux conversations informelles qu'on retrouve le plus de rires. L'examen de ces résultats montre que plus la parole est relâchée, plus la fréquence des rires augmente. On trouve également un grand nombre de bruits, en particulier dans *AixMapTask,* corpus pour lequel ils représentent le 4ème token le plus fréquent. Nous nous abstenons toutefois de tirer une conclusion relative au style de parole car la présence de bruits tels que les inspirations/expirations dépend de la position/qualité du micro. Nous constatons cependant qu'il peut être fréquent dans les données.

| **Corpus** | **Nombre de tokens** | **% de euh** | **% de rires** | **% de bruits** |
|---|---|---|---|---|
| *Europe* | 7 566 | 6,014% | 0,013% | 0,264% |
| *Typaloc* | 7 534 | 2,933% | 0,186% | 1,434% |
| *AixMapTask* | 37 979 | 2,285% | 0,635% | 2,607% |
| *CID* | 126 260 | 3,997% | 1,221% | 0,870% |
| *Cheese* | 16 829 | 2,793% | 2,246% | 0,434% |

TABLE 2 : Tokens et pourcentages que représentent les euh, rires et bruits

## 2.3 Contexte des euh, rires et bruits

La table 3 fait mention des contextes dans lesquels on retrouve les euh, rires et bruits. Effectivement, pour cette étude, dans laquelle nous nous intéressons plus particulièrement à la tâche de segmentation de la parole, il est utile de savoir dans quelle mesure le système automatique devra intervenir. La colonne 2 indique le pourcentage des euh, rires ou bruits qui sont entourés de silences, c'est-à-dire que l'événement constitue une IPU à lui seul. Dans ce cas, le système d'alignement forcé n'interviendra pas puisque l'événement a déjà été aligné par la segmentation en IPUs. Pour les colonnes 3 et 4, le système d'alignement forcé devra déterminer respectivement la frontière finale ou

initiale de l'événement. Enfin, la dernière colonne indique les cas où l'événement se trouve entouré d'autres tokens (soit des mots, soit un autre événement), donc le système d'alignement forcé aura à déterminer ses frontières initiales et finales. On observe que plus d'un tiers des rires et un bruit sur cinq sont entourés de silences. Quant au "euh", on peut dire qu'il ne se retrouve quasiment jamais isolé entre deux silences. Si l'on met en lien les tables 2 et 3, on en conclut que les euh, rires et bruits représentent un nombre important de tokens dans tous les corpus spontanés, mais dans des proportions différentes selon le style de parole.

| | **entouré de silences** | **au début d'une IPU** | **à la fin d'une IPU** | **au sein d'une IPU** |
|---|---|---|---|---|
| euh | 1,47% | 11,80% | 28,99% | 57,75% |
| rire | 34,65% | 19,07% | 29,09% | 17,19% |
| bruit | 21,19% | 28,40% | 11,84% | 38,58% |

TABLE 3 : Pourcentage des euh, rires et bruits en fonction de leur contexte gauche et droit

Pour terminer cette partie de l'étude des distributions des corpus, la table 4 indique le nombre et la proportion d'IPUs dans lesquelles on retrouve les euh, rires et bruits. Ces chiffres sont importants, en effet, puisque le système d'alignement forcé, qui opère séparément sur chacune des IPUs, utilise un algorithme d'optimisation global sur la séquence. A ce titre, il peut arriver qu'une erreur d'alignement affecte largement ses contextes droits et gauches. On voit ainsi que 20% à 36% des IPUs contiennent au moins un euh, un rire ou un bruit (dernière colonne).

| **Corpus** | **# total IPUs** | **IPUS avec "euh"** | **IPUs avec rire** | **IPUs avec bruit** | **IPUs avec au moins un euh/rire/bruit** |
|---|---|---|---|---|---|
| *Europe* | 875 | 35,88% | 0,11% | 2,29% | 35,89% |
| *Typaloc* | 522 | 28,25% | 2,68% | 14,94% | 35,82% |
| *AixMapTask* | 6126 | 12,16% | 3,67% | 13,52% | 20,60% |
| *CID* | 13631 | 27,32% | 10,25% | 7,52% | 32,14% |
| *Cheese* | 2675 | 14,62% | 12,45% | 2,73% | 21,16% |

TABLE 4 : Nombre d'IPUs dans lesquelles les euh, rires et bruits apparaissent

On constate également que ces événements se retrouvent très souvent dans les mêmes IPUs, puisque la dernière colonne est loin de représenter la somme des colonnes 2 à 4. Pour illustrer ce phénomène, nous avons extrait une IPU dans deux corpus :

- exemple de *Typaloc* : "donc *euh* des choses *euh* genre *euh* canard à l'orange des choses comme ça qui demandent *euh* une préparation un peu plus subtile une surveillance"
- exemple de *Cheese* : "tu vas avec ton père *euh* il repart avec mille chameaux à @"

# 3 Alignement forcé

## 3.1 Corpus de test et méthode d'évaluation

Un corpus de test a été manuellement phonétisé et aligné par un expert avec le logiciel Praat. Cette annotation a ensuite été vérifiée et éventuellement révisée par une seconde personne. Les fichiers de ce corpus ont été extrait aléatoirement du corpus CID. Il comprend 27 IPUs de 12 locuteurs différents, pour une durée totale de 141 secondes. En tout, 1833 labels devront être alignés par le système : 1791 phonèmes, 24 "euh", 5 rires, 4 bruits et 9 pauses courtes.

Les évaluations consistent à comparer la segmentation automatique à celle effectuée manuellement avec la mesure communément nommée "Unit Boundary Position Accuracy (UBPA)". Elle estime le pourcentage de frontières automatiques incluses dans une fenêtre d'une durée donnée autour des frontières manuelles correspondantes. C'est donc une mesure quantitative qui permet de situer globalement les performances d'un système, mais surtout, elle permet de comparer aisément et rapidement différents systèmes.

## 3.2 Résultats d'alignement avec la mesure UBPA

Pour réaliser cette étude, dans un premier temps, nous avons construit un modèle acoustique appris sur des données de parole lues. L'apprentissage a été réalisé à l'aide de la boîte à outils HTK (Young & Young, 1993) et de SPPAS, en suivant le tutoriel du site voxforge.org[4]. Un modèle HMM à 5 états du silence et de chacun des 31 phonèmes suivants ont ainsi été appris :

- voyelles : A/ E e 2 i O/ 9 u y
- voyelles nasalisées: a~ U~/ o~
- plosives : p t k b d g
- fricatives : f v s z S Z
- consonnes nasales : m n
- liquides : l R
- glides : H j w

pour lesquels A/ représente a ou A, O/ représente o ou O et U~/ représente e~ ou 9~, en SAMPA comme proposé par J.C. Wells[5]. Les euh, rires et bruits n'étant que peu présent voire absents des corpus lus, nous avons utilisé un modèle prototypique[6] pour chacun d'entre-eux. Le "euh" y est symbolisé par l'étiquette fp, le rire par lg et le bruit par gb. Par la suite, nous appellerons ce modèle "initial". Nous avons ensuite appris les modèles des euh, rires et bruits en suivant la même procédure d'apprentissage, que l'on a appliquée sur les corpus décrits dans la table 2. Ces derniers n'ont pas été utilisés pour l'apprentissage des modèles des phonèmes car le modèle appris à partir des données lues amène à des résultats significativement meilleurs (selon la mesure UPBA). Ces trois HMM ont alors été injectés dans le modèle initial, remplaçant ainsi les prototypes.

[4] A la différence du tutoriel, nous avons utilisé SPPAS pour normaliser, phonétiser et aligner les données pendant la phase d'apprentissage, en paramétrant SPPAS pour qu'il utilise *Julius* (Lee et al., 2001) plutôt que la commande *HVite* d'HTK lors de l'alignement. Nous avons utilisé la version 4.2.2 du "Open-Source Large Vocabulary CSR Engine Julius" : http://julius.osdn.jp

[5] http://www.phon.ucl.ac.uk/home/sampa/french.htm

[6] Modèle résultat de la commande *HCompV* d'HTK (version 3.4.1) : http://htk.eng.cam.ac.uk/

La table 5 indique les mesures UBPA avec une valeur de taille de fenêtre variant de 20ms à 80ms. Plusieurs modèles sont comparés : le modèle initial puis ce même modèle dans lequel on utilise soit le HMM prototype soit le HMM appris pour les euh, rires et bruits. Aucune modification n'est apportée aux autres HMM du modèle, et dans tous les cas, SPPAS est utilisé pour aligner (configuré pour appeler *Julius*). Ces résultats montrent que l'utilisation d'un modèle appris pour le bruit ne change pas les résultats. En revanche, l'introduction du HMM appris du rire améliore les performances, même dans ce corpus de test ne contient que 5 items. Enfin, l'introduction du HMM appris spécifiquement pour les euh augmente significativement la qualité de l'alignement.

| | 20ms | 30ms | 40ms | 50ms | 80ms |
|---|---|---|---|---|---|
| modèle initial<br>(euh, rires et bruits prototypes) | 84,91 | 92,16 | 94,32 | 95,45 | 97,02 |
| avec HMM des bruits appris<br>(euh et rires prototypes) | 84,75 | 92,10 | 94,37 | 95,51 | 97,08 |
| avec HMM des rires appris<br>(euh et bruits prototypes) | 85,13 | 92,48 | 94,75 | 95,94 | 97,67 |
| avec HMM des euh appris<br>(rires et bruits prototypes) | 86,05 | 93,67 | 96,00 | 97,08 | 98,48 |
| modèle final<br>(euh, rires et bruits appris) | 86,10 | 93,94 | 96,48 | 97,62 | 99,19 |

TABLE 5 : UBPA (%) du système d'alignement en utilisant différents modèles acoustiques

Pour compléter cette étude, nous ne pouvions pas ignorer le fait que les autres systèmes d'alignement forcé qui traitent la langue française utilisent la voyelle 2 pour aligner les "euh". Nous avons donc estimé les résultats en utilisant le modèle initial dans lequel le HMM du euh est remplacé par celui de la voyelle 2. Les mesures UBPA sont à 20ms : 85,67% ; à 30ms: 92,91% ; à 40ms : 95,19% ; à 50ms : 96,37% ; et à 80ms : 97,83%. Ainsi, l'utilisation de la voyelle 2 s'avère nettement plus judicieuse que l'utilisation du prototype, dans le cas où il ne serait pas possible de disposer d'un modèle spécifique. Nous avons néanmoins montré que l'apprentissage d'un modèle spécifique pour les "euh" amène à un meilleur résultat d'alignement significativement meilleur (avant dernière ligne du tableau) qu'en utilisant une voyelle acoustiquement proche.

## 3.3 Résultats qualitatifs des alignements

L'analyse qualitative des résultats présente un grand intérêt, notamment pour mieux comprendre et donc mieux appréhender les annotations obtenues. Nous nous sommes donc intéressés aux erreurs majeures, c'est-à-dire lorsque le système propose un désalignement qui dépasse 80ms, soit 15 cas dans notre corpus. Un point important à soulever concerne le fait que ces erreurs majeures se concentrent sur 5 IPUs seulement sur les 27 qui ont été alignées. La figure 1 illustre la séquence d'erreurs la plus importante du corpus, lors de l'alignement des tokens "na na na na na". Dans un segment de discours rapporté, le locuteur mentionne le fait que le discours continue mais ne présente pas d'intérêt pour le propos actuel. Bien qu'elle soit audible, il produit cette séquence de manière

assez hypo-articulée. Le système d'alignement échoue à déterminer les frontières entre les phonèmes n et A/ et 6 erreurs d'alignement sont ainsi observées sur cette suite de 12 phonèmes. La figure 2 illustre une autre cascade d'erreurs : le système ne détermine pas correctement le début du rire et lui assigne le phonème A/ du mot qui le précède, et cela affecte la suite des 4 phonèmes k-t-w-A/.

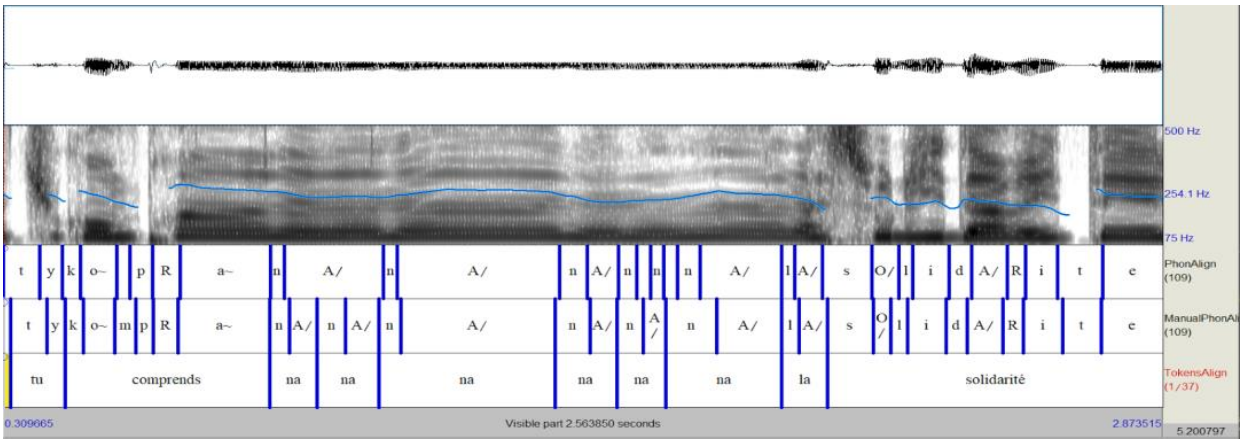


FIGURE 1 : Erreurs d'alignement sur la séquence de tokens "na na na na na na"

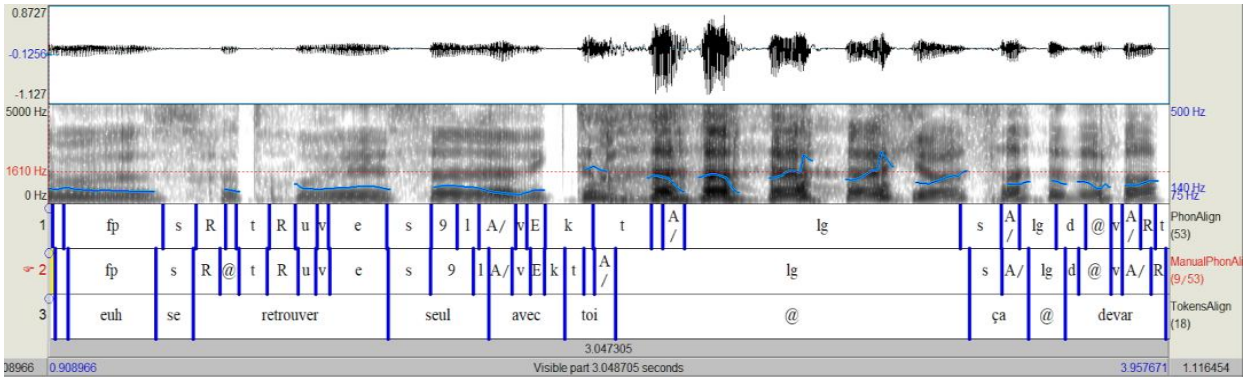


FIGURE 2 : Erreurs d'alignement sur la séquence de tokens "avec toi @"

# 4 Conclusion

Pour cette étude, nous avons rassemblé différents corpus de parole spontanée, en les sélectionnant de sorte qu'ils soient relativement homogènes dans leur forme (conditions d'enregistrement, transcription, etc). Ces corpus nous ont permis d'estimer les fréquences des tokens et de les comparer. Nous nous sommes intéressés en particulier aux *euh*, *rires* et *bruits*. Nous avons observé que le « euh » se retrouve fréquemment en parole spontanée, mais qu'il est nettement plus fréquent (6%) dans un débat politique que dans les conversations (2,3% à 4%). En revanche, plus la parole est relâchée, plus les rires sont présents ; ils peuvent en effet représenter jusqu'à 2,25% des tokens en conversation (sans consigne). On retrouve également une proportion non négligeable de bruits (produits par le locuteur) dans les différents corpus. Ces 3 items sont tellement fréquents en parole spontanée, qu'on les retrouve dans 20% à 36% des IPUs. En analysant leurs contextes gauche et droits, on constate qu'ils sont rarement isolés entre deux silences : seuls 1,47% des euh sont isolés. Ces résultats nous ont amené à évaluer l'impact de la prise en compte de ces 3 items dans la tâche d'alignement automatique : nous avons comparé l'utilisation d'un modèle acoustique qui inclut soit un HMM prototypique, soit un HMM appris. Nous en avons conclu qu'il n'était pas vraiment nécessaire d'apprendre un modèle de bruit. Des études plus poussées sur ce point nous permettraient

d'approfondir cet aspect. En revanche, la prise en compte des rires et surtout des euh, conduit à une nette amélioration des performances du système.

Les scripts et modules Python que nous avons développé pour l'apprentissage des modèles acoustiques ainsi que pour leur évaluation sont distribués sous licence GPL dans la version 1.9.4 de SPPAS et le modèle acoustique du français le sera dans la version 1.9.6.

# Références


ADDA-DECKER M., GENDROT C., NGUYEN N. (2008). Contributions du traitement automatique de la parole à l'étude des voyelles orales du français. Traitement Automatique des Langues, v. 49, n. 3, p. 13–46.

BERTRAND R., BLACHE P., ESPESSER R., FERRE G., MEUNIER C., PRIEGO-VALVERDE B., RAUZY S. (2008). Le CID — Corpus of Interactional Data — Annotation et Exploitation Multimodale de Parole Conversationnelle. Traitement Automatique des Langues, v. 49, n. 3.

BIGI B. (2015). SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech. *The Phonetician,* International Society of Phonetic Sciences, v. 111–112*,* p. 54–69.

GORISCH J., ASTÉSANO C., GURMAN BARD E., BIGI B., PRÉVOT L. (2014). Aix Map Task corpus: The French multimodal corpus of task-oriented dialogue. Proceedings of the 9th International conference on Language Resources and Evaluation, Reykjavik, Iceland, p. 2648–2652.

JOHNSON K. (2004). Massive reduction in conversational American English. In Spontaneous speech: Data and analysis. Proceedings of the 1st session of the 10th international symposium, Tokyo, Japan, p. 29–54.

LEE A., KAWAHARA T., SHIKANO K. (2001). Julius -- an open source real-time large vocabulary recognition engine. Proceedings of the European Conference on Speech Communication and Technology, Aalborg, Denmark, p. 1691–1694.

MEUNIER C., FOUGERON C., FREDOUILLE C., BIGI B., CREVIER-BUCHMAN L. ET AL. (2016). The TYPALOC Corpus: A Collection of Various Dysarthric Speech Recordings in Read and Spontaneous Styles. Proceedings of the 10th Language Resources and Evaluation Conference, Portorož, Slovenia. p. 4658–4665.

MEUNIER C. (2013). Phoneme deletion and fusion in conversational speech. Proceedings of the Experimental Approaches to Perception and Production of Language Variation, Copenhague, Danemark.

PORTES C. (2004). Prosodie et économie du discours : spécificité phonétique, écologie discursive et portée pragmatique de l'intonation d'implication. Université de Provence - Aix-Marseille I (PhD).

PRIEGO-VALVERDE B., BIGI B. (2016). Smiling behavior in humorous and non humorous conversations: a preliminary cross-cultural comparison between American English and French. International Society for Humor Studies Conference, Dublin, Ireland.

YOUNG S.J., YOUNG SJ. (1993). The HTK hidden Markov model toolkit: Design and philosophy. University of Cambridge, Department of Engineering.

# Gémination non-native en français d'apprenants italophones

Paolo Mairano[1], Fabian Santiago[2], Elisabeth Delais-Roussarie[3]
(1) LFSAG, Université de Turin, 10100, Turin, Italie.
(2) Université Paris VIII, SFL/LPP/LLF-CNRS, Paris, France
(3) UMR 6310-LLING, Université de Nantes, 44000 Nantes, France
paolo.mairano@unito.it, fabian.santiago-vargas@univ-paris8.fr,
elisabeth.delais-roussarie@univ-nantes.fr

## RÉSUMÉ

Le rôle de l'orthographe dans l'acquisition d'une L2 par des adultes a été un objet d'étude dans les recherches en psycholinguistique et en acquisition d'une L2. Plusieurs travaux récents ont montré que l'orthographe peut être à l'origine de contrastes phonologiques inexistants dans la langue cible, comme, par exemple, la production de consonnes pseudo-géminées en anglais L2 par des italophones. Dans cet article, nous étudions le cas de consonnes pseudo-géminées en correspondance des graphèmes <cc> vs <c> (ex. *immigrés* vs *imiter*) dans la parole lue d'apprenants italophones dans un corpus de français L2. Nous trouvons que les durées consonantiques associées à deux lettres sont plus longues que leurs contreparties associées à un seul symbole graphique. En revanche, les durées des voyelles qui précèdent ces consonnes ne subissent pas les modifications observées en italien L1, à savoir une réduction compensatoire. Différents facteurs pourraient expliquer ces indices divergents : il est possible que les apprenants essaient de s'écarter des réalisations pseudo-géminées qui rappellent leur L1 afin de se rapprocher du modèle du français natif.

## ABSTRACT

**Non-native gemination in L2 French by Italian natives.**
The role of orthography in adult L2 acquisition is well documented in the literature. Recent work has revealed that spelling can lead learners to produce phonological contrasts which do not exist in the target language, such as pseudo-geminate consonants in L2 English as produced by Italian speakers. In this contribution, we examine pseudo-geminate realizations in correspondence of <cc> vs <c> graphemes (e.g. *'immigrés'* vs *'imiter'*) as produced by Italian learners in a corpus of L2 French. We measure and compare the duration of the target consonant and the preceding vowel. The observed consonant durations in correspondence of <cc> spelling are longer than their <c> counterparts. Instead, the preceding vowels do not show the compensatory shortening which would be expected on the basis of L1 Italian gemination. Several factors may account for such divergent cues; among them, it is possible that learners try to depart from geminate realizations that sound like their L1 in the attempt to get closer to the French native model.

MOTS-CLÉS : prononciation, gémination, acquisition, français L2, effets de l'orthographe.
KEYWORDS: pronunciation, gemination, acquisition, L2 French, effects of orthography.


10.21437/JEP.2018-75

# 1 Introduction

L'apprentissage d'une langue étrangère est influencé chez les adultes par l'exposition à la langue écrite (au moins en partie, et souvent majoritairement). Ce phénomène a été bien décrit dans la littérature (Bassetti, Escudero & Hayes-Harb, 2015). Il n'est donc pas surprenant que l'orthographe ait des effets sur la prononciation des apprenants. D'une part, l'orthographe écrite peut favoriser la reconnaissance et la mémorisation des mots et des sons, car il peut fournir une représentation immuable de ces derniers (Erdener, Burnham, 2005). D'autre part, l'orthographe peut déclencher des prononciations divergentes de la norme de la langue cible, ce phénomène étant attesté par un nombre important de travaux de recherche dans ce domaine. Ce serait par exemple le cas pour les apprenants anglophones d'espagnol L2 qui produisent [v] à la place de [b] ou [β] (selon le contexte phonologique en espagnol L1) du fait de la présence du graphème <v> (Zampini 1994). Un deuxième exemple peut s'observer chez les apprenants italophones et japonophones de l'anglais L2, ces derniers produisant des consonnes longues ou courtes (désormais gémination non-native) en fonction de l'orthographe (ex. *Finnish* vs *finish*). La production de géminées non-natives résulterait de l'application d'une règle de correspondance graphème - phonème présente dans leur L1 (Bassetti, 2017 ; Bassetti et al., en révision ; Sokolovic-Perovic et al., en révision).

Ce dernier cas nous parait particulièrement intéressant, car il montre les effets que peut avoir l'orthographe dans la création des contrastes phonologiques non-natifs dans une L2. L'italien possède un contraste phonologique entre consonnes géminées et non-géminées très productif (ex. it. *sano* - *sanno*, fr. *sain* - *ils savent*). La gémination en italien est restreinte en position interne de mot, à l'exception du *raddoppiamento fonosintattico* (un cas de gémination post-lexicale), typique des variétés centrales et méridionales (Bertinetto & Loporcaro, 2005 ; cf., pour un cas différent, le berbère tachelhit, où la gémination peut également apparaitre en position initiale et finale (Ridouane 2007). La gémination en italien se manifeste principalement par un contraste de durée : les consonnes géminées seraient en moyenne deux fois plus longues que les consonnes non-géminées, avec un effet compensatoire sur la voyelle précédente qui se verrait raccourcie d'approximativement 25% (Esposito & Di Benedetto, 1995 ; Mattei & Di Benedetto, 2000). A l'opposé de l'italien, la gémination n'a pas un rôle lexical en français, où elle peut apparaitre essentiellement à la frontière de mots (ex. *avec quoi*) et après la chute d'un schwa (ex. *netteté*) (Hallé & Ridouane, 2011) ; des études récentes ont montré que même dans ces cas, les géminées ne sont pas produites de manière systématique (Meisenburg, 2006).

Nous présentons ici les résultats d'une étude sur corpus ayant un double objectif. Le premier est de confirmer si la présence de pseudo-gémination induite par l'orthographe est observée dans la production orale en français L2 par des italophones, à l'instar de l'étude de Bassetti (2017), cette dernière l'ayant observée dans la production des apprenants italiens en anglais L2. Le deuxième objectif est d'analyser comment se réalise la gémination non-native dans le signal de parole : nous mesurons donc non seulement la durée des consonnes cibles, mais également celle des voyelles précédente et suivante. A la différence des études menées par Bassetti et collègues (données obtenues par le biais de mots cibles à l'intérieur de phrases cadres ou de phrases isolées), nous avons opté pour un protocole légèrement plus écologique quoique moins contrôlé, où les participants lisent de petits textes, sans aucune répétition.

# 2 Méthode

Nous avons analysé les enregistrements de 25 étudiants italiens à l'université de Turin (Italie) suivant le protocole développé par Delais-Roussarie et al. (en préparation). Celui-ci a pour but la constitution d'un large corpus de données comparables d'apprenants de français L2 venant de plusieurs L1 (italien, allemand, suédois). Plusieurs tâches linguistiques ont été demandées aux participants, afin d'éliciter de la parole lue et de la parole semi-spontanée. L'analyse présentée dans cet article se base sur les données de parole lue, élicitées à travers un protocole spécifiquement conçu pour observer la pseudo-gémination en français L2, selon les modalités décrites dans le paragraphe 2.2.

## 2.1 Participants

Nous avons recruté 25 étudiants italophones à l'université de Turin qui poursuivaient des cours de français langue étrangère (niveaux B1, B2 et C1). Les locuteurs ont participé à l'expérience de manière rétribuée (6 euros). Tous les participants ont rempli un questionnaire contenant des informations sur leur âge (25.2, DS = 3.7), leur sexe (21 F et 4 H – ce déséquilibre reflétant la population d'étudiants de français à l'Université de Turin), leur région de provenance (16 participants : Piémont, 9 participants : autre), ainsi que sur le temps et les modalités d'acquisition du français (apprentissage scolaire ou informel, séjour en pays francophones, etc.). L'âge du premier contact avec la langue française se situait en moyenne à 12 ans, mais avec des différences importantes parmi les participants (min = 6 ans, max = 21 ans). 24 sur 25 participants avaient déjà été au moins une fois dans un pays francophone et 5 d'entre eux avaient réalisé des échanges universitaires *Erasmus* en France.

## 2.2 Tâche

Les 25 participants ont été enregistrés dans une chambre sourde dans les locaux de l'Université de Turin alors qu'ils lisaient à voix haute huit textes en français, ces derniers correspondant à des dialogues simulant des scènes de la vie quotidienne et à des histoires brèves et simples. Ces textes contenaient 48 mots cibles qui ont été insérés loin de toute frontière prosodique prévisible (c'est-à-dire en évitant la fin de syntagme intonatif ou de proposition). Cela visait à éviter des allongements de la durée qui refléteraient la structuration prosodique du français (notamment les allongements finaux).

| | **Attaque de 2ème syllabe** | **Coda de dernière syllabe** |
|---|---|---|
| [p] | *capacité - appareil*<br>*proportions - opportun*<br>*proposer - opposer* | *râpes - nappe*<br>*tape - frappes*<br>*étape - échappe* |
| [t] | *catalan - attaché*<br>*latitude - attitude*<br>*pratiquer - attirer* | *pates - pattes*<br>*vite - quitte*<br>*achète - brochette* |
| [m] | *promettons - commerçants*<br>*samaritain - grammaticaux*<br>*imiter - immigrés* | *dame - femme*<br>*lame - flamme*<br>*deuxième - dilemme* |

| | | |
|---|---|---|
| [n] | *animaux - anniversaire*<br>*phonologie - connotation*<br>*inoffensive - innovateur* | *gêne - chienne*<br>*laine - Rennes*<br>*semaine - antennes* |

TABLE 1 : Les 48 mots cibles insérés dans les textes.

Les 48 mots cibles (cf. table 1) constituent 24 paires de mots. A l'intérieur de chaque paire, les deux mots contiennent une même consonne ([p], [t], [m] ou [n]) écrite avec une (<c>) vs deux (<cc>) lettres. Ces consonnes forment des oppositions de gémination très productives en italien. De même, elles représentent deux modes d'articulation (occlusif, nasal), deux lieux d'articulation (bilabial, alvéolaire) et deux types de mécanismes laryngés : sourdes ([p], [t]) vs sonores ([m], [n]). Les consonnes cibles associées à un graphème <c> sont toujours précédées et suivies des mêmes voyelles que celles de leurs contreparties associées à un graphème <cc> (ex. : *pr**ati**quer-**atti**rer*). Nous avons essayé dans la mesure du possible d'avoir le même nombre de syllabes pour chaque item d'une même paire (par exemple : *imiter – immigrés*), et lorsqu'il était difficile d'y parvenir, nous avons tenté de nous en approcher (par exemple *animaux-anniversaire*). En outre, nous avons examiné les effets dans deux positions différentes dans le mot : (a) en tête de la deuxième syllabe dans des mots trisyllabiques ou tétrasyllabiques, donc précédant d'au moins une syllabe l'accent final du français (ex. : *imiter-immigrés*) ; (b) finale de mot (en coda de la dernière syllabe), suivant immédiatement la voyelle portant l'accent final (ex. : *râpes-nappes*).

En ce qui concerne les mots cibles avec <cc> en position finale, nous remarquons que la prononciation des italophones pourrait être affectée par deux phénomènes contrastants provenant de leur L1 : d'une part, les restrictions phonosyntaxiques de l'italien empêchent la gémination en fin de mot, ce qui devrait donc défavoriser l'allongement des consonnes cibles dans cette position. D'autre part, les effets de la gémination en italien sont plus importants en position accentuée (Payne, 2005), ce qui devrait donc favoriser l'allongement de ces consonnes. Au vu de ces éléments, nous avons examiné dans quelle mesure ces patrons de l'italien L1 étaient transposés en français L2 en contrôlant des positions potentiellement candidates à déclencher un tel transfert.

# 3 Résultats et analyse

Les enregistrements ont été numérisés au format wav (44 kHz). La transcription phonétique des textes a été alignée sur le signal sous Praat avec le plug-in *EasyAlign* (Goldman, 2011). La transcription et la segmentation ont ensuite été vérifiées manuellement. Les durées des consonnes cibles et des voyelles adjacentes ont été extraites via un script Praat développé ad hoc, elles ont ensuite été sauvegardées au format .csv et importées sous R pour l'analyse statistique. Les mots cibles comportant une hésitation ou une faute de lecture ont été exclus.

## 3.1 Durées des consonnes cibles

La durée consonantique étant le corrélat principal de la gémination en italien (Esposito & Di Benedetto, 1999), nous avons comparé les durées moyennes des consonnes cibles. Nous avons obtenu 1177 consonnes cibles dont trois ont été exclues de l'analyse, leur durée dépassant 250 ms. La table 2 montre que les durées des consonnes cibles <cc> sont en moyenne plus longues que les durées des consonnes cibles <c>, avec un ratio cc:c autour de ~1.25. Cela reflète partiellement les résultats obtenus pour l'anglais L2 par des locuteurs italiens et japonais (v. section 2), ou le ratio

moyen était de ~1.5. Une analyse supplémentaire par locuteur révèle que le ratio cc:c moyenné sur les mots cibles oscille entre 1.08 et 1.4, avec une médiane de 1.23.

La figure 1 présente les durées des consonnes selon leur position dans le mot. Ces figures suggèrent que le nombre de lettres (<c> vs <cc>) et la position dans le mot affectent la durée des consonnes. Afin de valider ces observations, nous avons construit un modèle linéaire à effets mixtes où nous avons entré la durée consonantique comme variable dépendante ; le graphème de la cible (<c> vs <cc>), la consonne ([p] vs [t] vs [m] vs [n]) et la position dans le mot (deuxième syllabe vs dernière syllabe) comme effets fixes ; et les participants et les mots comme effets aléatoires. Les résultats de ce modèle statistique auprès d'un test du rapport de vraisemblance (*likelihood ratio test*) ont confirmé cette observation. En ce qui concerne l'effet du graphème, <cc> correspond à une augmentation de +25 ms de la durée de la consonne par rapport à <c> ($\chi^2(1) = 25.76$, $p < .001$), après avoir contrôlé toute autre variable. Les résultats montrent également que la durée des consonnes est affectée par le type de consonne ($\chi^2(3) = 17.32$, $p < .001$) et par sa position à l'intérieur du mot ($\chi^2(1) = 10.05$, $p = .002$). Enfin les résultats montrent que la durée des consonnes est affectée également par l'interaction entre le type de consonne et la position dans le mot ($\chi^2(3) = 7.79$, $p = .05$), mais pas par l'interaction de ces derniers facteurs avec l'effet du graphème ($p > .05$). Celui-ci est donc constant pour les 4 consonnes et dans les 2 positions.

| consonne | durée <c> | durée <cc> | ratio cc:c |
|---|---|---|---|
| [m] | 74 (SD=23) | 95 (SD=26) | 1.28 |
| [n] | 69 (SD=24) | 81 (SD=28) | 1.18 |
| [p] | 86 (SD=25) | 105 (SD=30) | 1.22 |
| [t] | 80 (SD=25) | 106 (SD=39) | 1.31 |

TABLE 2 : Durées moyennes en ms pour les consonnes cibles.

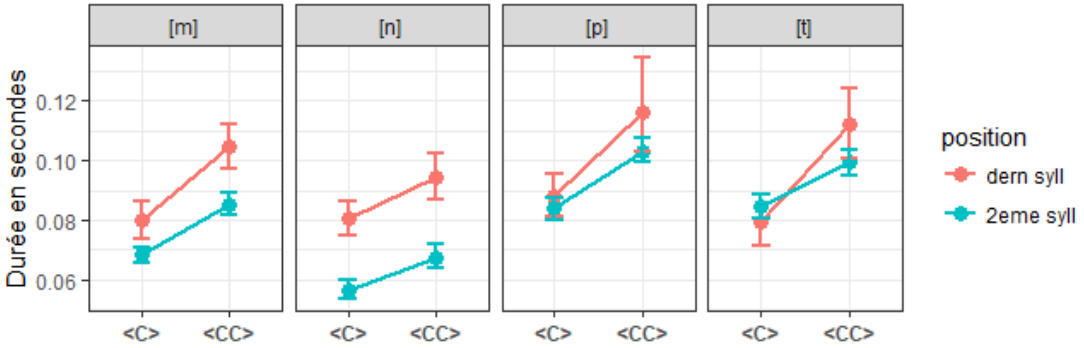



FIGURE 1 : Durées moyennes des consonnes cibles, selon leur position dans le mot.

## 3.2 Durées des voyelles précédant la consonne cible

En italien la durée de la voyelle qui précède une consonne géminée est, elle aussi, affectée : elle est réduite de 25% (Esposito & Di Benedetto, 1999). Aussi avons-nous également comparé les durées des 1177 voyelles précédant la consonne cible. Les valeurs dans la table 3 illustrent que, de manière

inattendue, les durées des voyelles précédant les consonnes cibles avec un graphème <cc> sont allongées au lieu d'être réduites. L'analyse par locuteur révèle que les ratios d'allongement vocalique oscillent entre 1.07 et 1.35, avec une médiane de 1.19. Les effets de la position dans le mot sont visibles dans la figure 2.

Suivant la même procédure que pour les durées consonantiques, nous avons construit un modèle linéaire à effets mixtes où nous avons entré la durée de la voyelle comme variable dépendante ; le graphème de la cible (<c> vs <cc>), la voyelle ([i] vs [ɛ] vs [a] vs [ɔ]) et la position dans le mot comme effets fixes ; et les participants et les mots comme effets aléatoires. Les résultats de ce modèle statistique auprès d'un test du rapport de vraisemblance (*likelihood ratio test*) ont révélé des effets différents de notre hypothèse. En ce qui concerne l'effet du graphème, <cc> correspond à une augmentation moyenne de +16 ms de la durée de la voyelle précédant par rapport à <c> ($\chi^2(1) = 9.36$, $p = .002$), après avoir contrôlé toute autre variable. Les résultats montrent également que la durée des voyelles n'est pas affectée par le type de voyelle ($\chi^2(3) = 1.73$, $p = .17$), mais par sa position dans le mot ($\chi^2(1) = 21.10$, $p < .001$), ce dernier effet reflétant l'allongement dû à l'accent primaire en français L1, ainsi qu'en italien L1. Enfin les résultats montrent que la durée des voyelles est affectée également par l'interaction entre le type de consonne et la position dans le mot ($\chi^2(1) = 6.274$, $p = .012$), mais pas par l'interaction de ces derniers facteurs avec l'effet du graphème ($p > .05$). Celui-ci est donc constant pour les 4 consonnes et dans les 2 positions.

| Consonne | durée v_<c> | durée v_<cc> | ratio v_cc:v_c |
|---|---|---|---|
| [m] | 71 (SD=35) | 92 (SD=41) | 1.29 |
| [n] | 76 (SD=31) | 85 (SD=39) | 1.12 |
| [p] | 75 (SD=24) | 87 (SD=26) | 1.16 |
| [t] | 72 (SD=21) | 88 (SD=32) | 1.22 |

TABLE 3 : Durées moyennes en ms pour les voyelles précédant les consonnes cibles.

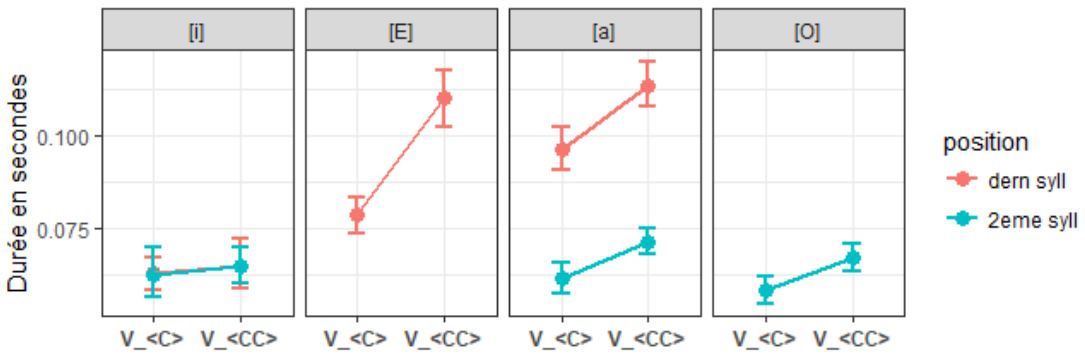



FIGURE 2 : Durées moyennes des voyelles précédant la consonne cible, selon leur position dans le mot.

En outre, nous avons testé les effets du graphème <c> vs <cc> sur la durée de la voyelle suivante (exclusivement pour les cibles en deuxième syllabe, puisque, évidemment, les consonnes cibles en position finale de mot ne sont généralement pas suivies d'une voyelle, cela dépendant du mot qui suit). Aucun effet significatif du graphème n'a été trouvé ($\chi^2(1) = 0.09$, $p = .76$).

# 4 Discussion et conclusion

Les résultats montrent que les locuteurs italophones modifient les durées consonantiques en conformité avec notre hypothèse du départ : la condition <c> vs <cc> semble entraîner la production de pseudo-géminées en français L2. En effet, nous avons montré que le choix du graphème représentant la consonne cible a un effet significatif sur la durée de celle-ci et de la voyelle qui la précède : la consonne comme la voyelle se rallongent dans le cas du graphème <cc>. L'allongement de la consonne cible indique que les apprenants appliquent une règle de correspondance graphème-phonème provenant de leur L1, de manière similaire à ce que font en anglais L2 les apprenants italophones (Bassetti, 2017) et japonophones (Sokolovic-Perovic et al., en révision). Nos données montrent que les ratios d'allongement de la consonne cible sous la condition <cc> sont en moyenne de ~1.2. Ils sont donc plus réduits que les ratios reportés dans la littérature pour l'italien L1 (~2 selon Esposito & Di Benedetto, 1999), mais aussi pour l'anglais L2 des italophones (autours de ~1.5 d'après les études citées précédemment). Plusieurs facteurs pourraient contribuer à expliquer ces différences.

Les différences de ratios dans la L2 par rapport à la L1 sont probablement le résultat d'un système hybride, une interlangue qui s'éloigne de la L1 et s'approche partiellement du français natif (la grande majorité de nos participants ayant été dans un pays francophone au moins une fois et ayant en général un bon niveau en français). Des phénomènes similaires ont été retrouvés par exemple dans les valeurs de VOT (*Voice Onset Time*) observées par Flege et al. (1995) : les locuteurs italophones de l'anglais L2 produisent des VOTs avec des valeurs intermédiaires entre celles reportées pour l'italien L1 et pour l'anglais L1. En revanche, la différence entre les ratios observés dans cette étude et les ratios observés dans les études sur l'anglais L2 des italophones et des japonophones s'explique probablement par des différences dans le protocole expérimental (lecture de textes continus vs lecture répétée de phrases) et/ou, également par le profil des participants (dans notre cas, des étudiants universitaires de français, donc fortement motivés) et par leur provenance géographique (le nord vs le centre de l'Italie).

Si l'allongement de la consonne représentée par la graphie <cc> confirme nos prédictions, l'allongement de la voyelle ne correspond pas à notre hypothèse initiale. D'après les études consacrées à l'italien central, la présence d'une consonne géminée aurait pour effet la réduction de la voyelle précédente dans le cas des plosives (Esposito & Di Benedetto, 1999) mais aussi des nasales (Mattei, Di Benedetto, 2000). Ce n'est pas ce que nous avons observé dans notre étude. De fait, les locuteurs ont allongé la voyelle au lieu de la raccourcir. On pourrait émettre plusieurs hypothèses pour en rendre compte. Tout d'abord, on pourrait penser que l'abrégement de la voyelle qui précède la géminée ne caractérise pas les géminées dans la L1 de nos locuteurs (de provenance du Piémont dans la plupart des cas). Une autre explication possible serait que la réalisation de la pseudo-gémination en L2 affecte un domaine plus large si bien que l'allongement touche la séquence VC dans son entier. Cette modalité de (pseudo-)gémination serait similaire à la modalité de gémination du japonais L1 où la voyelle s'allonge parallèlement à la consonne (Idemaru & Guion-Anderson, 2010). Pour finir, une autre explication est envisageable : les apprenants essaieraient de dissimuler les consonnes géminées en les rendant perceptivement moins saillantes (le raccourcissement de la voyelle est un indice perceptif secondaire de la gémination d'après Esposito & Di Benedetto, 1999) dans le but de se rapprocher du français natif. Un phénomène analogue a été relevé dans la littérature pour les locuteurs de l'anglais d'Écosse ; ces derniers essaieraient de dissimuler la prononciation des /r/ postvocaliques de l'anglais dans l'effort de se rapprocher du modèle plus prestigieux du sud de l'Angleterre (Lawson, Scobbie & Stuart-Smith, 2014). Bien qu'une réponse définitive ne soit pas possible à ce jour, l'analyse des données de l'italien L1 de ces locuteurs (lesquelles sont déjà récoltées) pourrait nous éclairer à cet égard.

# Remerciements

L'enregistrement des données a été possible grâce au soutien financier de l'UMR 7110-LLF, grâce auquel les participants ont pu être rétribués. En outre, ce travail a été financé par le projet ANR Labex EFL (*Empirical Foundations of Linguistics*) et a été mené dans l'opération « Ressources langagières : données, lexiques, corpus, outils » (Axe 6). Nous souhaitons remercier Valentina De Iacovo (Université de Turin) pour le support logistique qu'elle nous a offert pendant les enregistrements et pour l'aide fournie dans la phase de recrutement des participants.

# Références

BASSETTI B. (2017). Orthography affects second language speech: double letters and geminate production in English. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 43(11), 1835-1842.

BASSETTI B., ESCUDERO P., HAYES-HARB R. (2015). Second language phonology at the interface between acoustic and orthographic input. *Applied Psycholinguistics* 36(1), 1-6.

BASSETTI B., SOKOLOVIC-PEROVIC M., MAIRANO P., CERNI T. (en révision). Orthography-induced length contrasts in the second language phonological systems of experienced speakers of English as a Second Language: evidence from minimal pairs, en révision.

BERTINETTO P.M., LOPORCARO M. (2005). The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome. *Journal of the International Phonetic Association,* 35(2), 131-151.

DELAIS-ROUSSARIE E., KUPISCH T., MAIRANO P., SANTIAGO F., SPLENDIDO F. (en préparation). ProSeg: A comparable corpus of spoken L2 French, en préparation.

ERDENER V.D., BURNHAM D.K. (2005). The role of audiovisual speech and orthographic information in nonnative speech production. *Language Learning,* 55(2), 191-228.

ESPOSITO A., DI BENEDETTO M.G. (2000). Acoustical and perceptual study of gemination in Italian stops. *The Journal of the Acoustic Society of America*., 104(6), 2051-2062.

FLEGE J.E., MUNRO M.J, MACKAY I.R. (1995). Effects of age of second-language learning on the production of English consonants. *Speech Communication*, 16(1), 1-26.

GOLDMAN B. (2011). EasyAlign: an automatic phonetic alignment tool under Praat. Actes de *Interspeech2011*.

HALLÉ P., RIDOUANE R. (2011). French listeners' deafness to Tashlhiyt Berber /bi/-/bbi/. Actes de *ICPhS XVII*, 811-814.

IDEMARU K., GUION-ANDERSON S. (2010). Relational timing in the production and perception of Japanese singleton and geminate stops. *Phonetica*, 67(1-2), 25-46.

LAWSON E., SCOBBIE, J.M., STUART-SMITH J. (2014). A socio-articulatory study of Scottish rhoticity. *Sociolinguistics in Scotland*, Palgrave Macmillan, London, 53-78.

MATTEI M., DI BENEDETTO M.G. (2000). Acoustic analysis of singleton and geminate nasals in Italian. *The European Journal of Language and Speech*, 1-11.

PAYNE E. (2005). Phonetic variation in Italian consonant gemination. *Journal of the International Phonetic Association* 35(2), 153-181.

RIDOUANE R. (2007). Gemination in Tashlhiyt Berber: an acoustic and articulatory study. *Journal of the International Phonetic Association* 37(2), 119-142.

SOKOLOVIC M., DILLON S., BASSETTI B. (en révision). Effects of orthographic forms on phonology in Japanese speakers of English as a Second Language, en révision.

ZAMPINI M.L. (1994). The role of native language transfer and task formality in the acquisition of Spanish spirantization. *Hispania*, 470-481.

# La distinction entre les paraphasies phonologiques et phonétiques dans l'aphasie : Étude acoustique des productions de 6 patients aphasiques

Clémence Verhaegen[1], Véronique Delvaux[1,2], Kathy Huet[1], Sophie Fagniart[1], Myriam Piccaluga[1], Bernard Harmegnies[1]
(1) Unité de Métrologie et Sciences du Langage, Université de Mons, Belgique
(2) Fond National de la Recherche Scientifique, Belgique
clemence.verhaegen@umons.ac.be

## RESUME

Notre objectif est de contribuer à la description des atteintes phonologiques et phonétiques dans l'aphasie. En effet, la distinction entre ces atteintes reste actuellement débattue, tant sur le plan méthodologique qu'épistémologique. Nous avons mené une étude de cas multiples auprès de 6 patients aphasiques. L'originalité de notre étude réside dans le fait que nous avons utilisé des techniques issues à la fois de la neuropsychologie du langage - en présentant des tâches classiques de dénomination d'images et de répétition-, mais également de la phonétique-en procédant à des analyses du VOT des contoïdes plosives dans une tâche de répétition de non-mots. Les résultats montrent la présence de troubles mixtes chez les patients et des profils d'erreurs qui diffèrent des hypothèses classiquement présentées dans la littérature. La réalité d'une distinction entre les erreurs phonologiques et phonétiques, de même que les théories reliées sont discutées.

## ABSTRACT

**The distinction between phonological and phonetic paraphasias in aphasia: An acoustic study of the speech outputs of six aphasic patients**

The aim of this study is to contribute to the description of phonological and phonetic impairment in aphasia. Indeed, the distinction between both deficits is actually debated, both methodologically and theoretically. Our study consisted in a multiple case study of 6 aphasic patients. The originality of our study lies in the fact that we used methods and techniques borrowed both the neuropsycholinguistics - by presenting picture naming and repetition tasks-, and phonetics-by conducting acoustic analyses of the VOT of plosive contoids on a nonwords repetition task. The results showed mixed impairment in our aphasic patients and error patterns that differ from the classic hypothesis presented in the literature. The distinction reality between phonological and phonetic errors, as well as the underlying theories are discussed.

MOTS-CLÉS : erreurs phonétiques, erreurs phonologiques, aphasie, analyse acoustique, VOT
KEYWORDS: phonetic errors, phonological errors, aphasia, acoustic analysis, VOT

# 1 Introduction

Historiquement, les aphasiologues étudiant les troubles de la production du langage ont créé une distinction, profondément ancrée dans la littérature ainsi que dans la clinique du langage, entre des troubles langagiers affectant la sélection des phonèmes au sein du système phonologique, identifiés comme des *atteintes phonologiques*, et des troubles, qualifiés de plus moteurs, affectant la programmation et l'exécution motrice des mouvements nécessaires aux réalisations des phonèmes,


10.21437/JEP.2018-76

identifiés comme des *atteintes phonétiques*. L'argument majeur en faveur de cette distinction a été l'observation de deux types d'erreurs distincts, attribuées respectivement à des atteintes phonologiques ou phonétiques : les erreurs ou paraphasies « phonologiques » se caractérisant par des ajouts, omissions, permutations ou substitutions de phonèmes au sein du mot, sans altération de la réalisation articulatoire, et les erreurs dites « phonétiques » se caractérisant par des distorsions de la réalisation des phonèmes (Galluzzi, Bureca, Guariglia, & Romani, 2015; Pillon & de Partz, 2014; Romani, Olson, Semenza, & Granà, 2002). Ce contraste a également été pris en compte dans les modèles classiques de la production du langage. Pour beaucoup d'auteurs, la planification des mots à produire s'effectue en deux étapes distinctes. D'abord, le niveau phonologique assure la planification de la forme abstraite des mots à produire; ensuite, les programmes moteurs articulatoires sont spécifiés au niveau phonétique (Dell, Schwartz, Martin, Saffran, & Gagnon, 1997; Indefrey, 2011; Levelt, 1999; Rapp & Goldrick, 2000).

Cependant, bien que la distinction entre les erreurs phonologiques et phonétiques soit considérée comme un fait relativement établi, lorsque celles-ci sont examinées avec précision, leur description reste soumise à de nombreux problèmes méthodologiques et épistémologiques. Une des principales difficultés est que les erreurs sont loin d'être toutes aisément détectables et validement analysables à l'écoute et, dès lors, sont exposées à des biais d'analyse lorsque l'investigation est essentiellement perceptuelle, ce qui est très fréquemment le cas (Nespoulous, Baqué, Rosas, Marczyk, & Estrada, 2013). Des tentatives d'objectivation par des études acoustiques ont été proposées dans la littérature. La plupart de ces travaux se sont centrés sur les contoïdes plosives et ont, pour la majorité, recouru à l'analyse du Voice Onset Time (VOT), qui mesure le délai entre le relâchement de l'occlusion supra-glottique et l'apparition des vibrations laryngées (Lisker & Abramson, 1964). Le VOT est le principal paramètre de l'opposition entre les occlusives sourdes et sonores dans un grand nombre de langues (Cho & Ladefoged, 1999) et constitue un indice important du contrôle des relations temporelles et de la coordination entre les gestes glottiques et supra-glottiques. Il est par conséquent parfaitement approprié pour l'étude des troubles de la planification et de l'exécution motrice des sons de parole chez les patients aphasiques. Les auteurs ont montré la valeur ajoutée de la démarche d'analyse phonétique en établissant (chez des locuteurs anglophones surtout) l'existence de profils distincts au niveau acoustique chez les patients présentant des troubles phonologiques et phonétiques. Ainsi, chez les patients diagnostiqués à trouble phonétique, on a montré une plus grande variabilité du VOT, incluant notamment difficultés de tenue du voisement, ainsi que de recouvrements inter-catégoriels. Tandis que chez les patients qui présentent des troubles phonologiques, on a montré des voisements et dévoisements de plosives, sans tendance préférentielle pour un phénomène plutôt qu'un autre, ainsi que des changements de lieux d'articulation. L'ensemble des valeurs des VOT pour les plosives examinées restant toujours proches des valeurs prototypiques observées pour ces deux catégories phonologiques dans la langue du participant (Baqué, Marczyk, Rosas, & Estrada, 2015; Blumstein, Cooper, Goodglass, Statlender, & Gottlieb, 1980; Kurowski & Blumstein, 2016; Nespoulous et al., 2013; Ryalls, Provost, & Arsenault, 1995). Les études restent néanmoins peu nombreuses, même en langue anglaise, extrêmement rares en langue française, et productrices de constats variables d'une étude à l'autre. En outre, ces études sont rarement reliées à d'autres indices et analyses, telles que des analyses neuropsycholinguistiques (ex. analyses des variables psycholinguistiques), qui permettent de compléter le profil du patient et d'apporter quelques lumières à des résultats posant question. Enfin, sur le plan épistémologique, il convient de souligner qu'un grand nombre d'études montrent une tendance à classer les patients a priori dans une catégorie de trouble (phonologique ou phonétique) en se basant sur la localisation de la lésion (temporale ou frontale) ou le type d'aphasie (fluente ou non fluente, aphasie de Broca vs. aphasie de conduction) (Blumstein et al., 1980; Galluzzi et al., 2015; Kurowski & Blumstein, 2016; Nespoulous et al., 2013; Romani et al., 2002; Ryalls et al., 1995). Ce type de raisonnement confine dès lors à la tautologie, puisque l'on recherche dans les

comportements des marques tantôt supposées phonétiques, tantôt phonologiques d'un trouble pré-étiqueté, tout en attendant que l'analyse des productions confirme la catégorisation clinique.

Dès lors, la distinction entre les erreurs phonologiques et phonétiques reste peu précise à l'heure actuelle. Or, la différentiation entre ces erreurs est importante puisqu'elle conditionne la démarche de rééducation langagière qui sera mise en place pour le patient cérébrolésé. Notre étude s'inscrit dans ce cadre épistémologique. Notre objectif est de contribuer à la description des atteintes phonologiques et phonétiques dans l'aphasie. Nous avons mené une étude auprès de 6 patients aphasiques de langue maternelle française. 4 patients présentaient une aphasie non fluente et 2 patients une aphasie fluente. L'originalité de notre étude réside dans le fait que nous avons utilisé des techniques issues à la fois de la neuropsychologie du langage mais également de la phonétique. Ainsi, nous avons présenté des tâches de production langagières classiques de dénomination d'images et de répétition de syllabes et de mots, dans lesquelles nous avons analysé les effets de variables psycholinguistiques telles que la longueur (dénomination et répétition) et la complexité articulatoire (répétition)[1], connues pour affecter les niveaux phonologique et phonétique dans la littérature en aphasiologie (Romani et al., 2002). Ensuite, en vue de caractériser ces troubles affectant la production de la parole chez ces patients, nous avons réalisé une analyse acoustique de leurs productions dans une épreuve de répétition de non-mots. Nous nous sommes principalement centrés sur l'analyse du VOT des consonnes occlusives voisées et non voisées. Conformément à la littérature existante, nous nous attendions à ce qu'un trouble phonétique soit marqué par la présence d'effets de longueur en dénomination et répétition et d'un effet de complexité articulatoire en répétition. Dans la tâche de répétition de non-mots destinée à étudier le VOT, nous nous attendions à ce que les patients montrent des difficultés de tenue du voisement lors de la production de voisées. En effet, les occlusives voisées du français présentent un VOT négatif long qui demande une coordination précise entre les articulateurs, fréquemment atteinte en cas de trouble phonétique. Nous nous attendions dès lors à la présence de VOT négatifs moyens plus courts que les participants contrôles pour les voisées et/ou un nombre plus important de dévoisements complets, ainsi qu'une plus grande variabilité dans les productions. En cas de trouble phonologique, nous nous attendions à des effets de longueur en dénomination et répétition mais pas d'effet de complexité articulatoire. Dans la tâche de répétition de non-mots, nous nous attendions à la présence de voisements et de dévoisements, dont les valeurs des VOT resteraient cependant dans les normes des réalisations observées en langue française, ainsi que des substitutions phonologiques (changements de lieu ou de mode d'articulation). Enfin, conformément à la littérature, nous devrions observer des manifestations de troubles phonétiques chez les patients non fluents et des manifestations de troubles phonologiques chez les patients fluents. Cependant, les patients n'ont pas été pré-classifiés dans une catégorie de trouble, afin de rompre avec le raisonnement circulaire décrit ci-dessus. Nous avons dès lors réalisé des analyses de cas uniques et non des études de groupes en fonction de leur type d'aphasie.

# 2 Participants

Six patients aphasiques, IJ, CL, TM, DG, BD et DM, de langue maternelle française ont participé à la présente étude. 4 patients (IJ, CL, TM, DG) ont été diagnostiqués par des spécialistes du langage, comme présentant une aphasie de Broca, non fluente, le patient BD une aphasie de Wernicke, fluente, et DM une aphasie de conduction, fluente également. Tous les patients présentaient une vue non altérée ou corrigée et pas d'atteinte auditive. Les patients ne présentaient

[1] La notion de complexité articulatoire varie dans la littérature (e.g., Romani et al., 2002). Les auteurs de cette tâche ont choisi de ne se concentrer que sur la présence ou non de groupes consonantiques, qui sont fréquemment altérés chez les patients aphasiques, apraxiques ou dysarthriques, qui présentent des troubles phonétiques (Nespoulous et al., 2013 ; Romani et al., 2002)

pas d'altération importante de la compréhension du langage, évaluée à l'aide de tâches de désignations de mots (*Examen Long du Langage*, UCL-ULg) ou de phrases (*Montréal-Toulouse*, Joanette et al., 1998). En outre, tous les patients montraient une atteinte de la MCT, et les patients IJ, BD, CL, TM également une altération des fonctions exécutives de mise à jour, flexibilité et inhibition. La Table 1 présente leurs principales caractéristiques. Les performances des patients ont été comparées à des participants contrôles, appariés en âge (*N*=8 pour les groupes de 46-54 et 70-79 ans et 10 pour le groupe de 50-59 ans).

| Patient | Âge | Genre | Type d'aphasie | Temps post-onset | Etiologie | Lésion | Groupe contrôle |
|---|---|---|---|---|---|---|---|
| IJ | 44 | F | Broca | 18 mois | AVC | Fronto-pariétale | 46-54 ans |
| CL | 65 | M | Broca | 2 ans | AVC | Fronto-pariétale | 60-69 ans |
| TM | 62 | M | Broca | 11 ans | AVC | Fronto-temporale | 60-69 ans |
| DG | 74 | M | Broca | 2 ans | AVC | Fronto-temporale | 70-79 ans |
| DM | 62 | M | Conduction | 10 ans | Trauma crânien | Temporale | 60-69 ans |
| BD | 72 | M | Wernicke | 18 mois | AVC | Pariétale | 70-79 ans |

TABLE 1 : Résumé des informations relatives aux participants de notre étude.

# 3 Méthodes et résultats

La présentation des tâches et des résultats associés est séparée en deux parties. Dans un premier temps, nous présentons les tâches langagières classiques et les résultats obtenus dans ces tâches, puis ceux liés à la tâche de répétition de non-mots, créée afin d'étudier le VOT. Les patients ont été évalués individuellement à leur domicile dans un local calme. Nous leur avons présenté les tâches sur 3 jours différents afin de ne pas les fatiguer. Chaque séance durait entre 45 et 60 minutes. L'ordre des tâches était le suivant : Jour 1 : (1) Anamnèse, (2) Dénomination d'images (40 premiers items), (3) Répétition de non-mots VOT (18 premiers items), (4) Désignation de phrases ; Jour 2 : (1) Dénomination d'images (40 derniers items), (2) Répétition de non-mots VOT (fin), (3) Désignation de mots ; Jour 3 : (1) Évaluation des fonctions exécutives, (2) Audiométrie tonale.

| | Dénomination | | | Répétition de syllabes | | | Répétition de mots | | |
|---|---|---|---|---|---|---|---|---|---|
| | Score (%) | Erreurs | Effet longueur | Score (%) | Erreurs | Effets variables | Score (%) | Erreurs | Effets variables |
| IJ | 50.00* | Phon. | non | 46.67* | Phon. Simpl.gp. | Comp. syll | 50.00* | Phon. | Comp. syll |
| CL | 73.75 | Phon. | non | 66.67 | Phon. Simpl.gp. | Comp. syll | 83.33* | Phon. | / |
| TM | 66.25* | Phon. | oui | 86.67 | Phon. | / | 77.78* | Phon. Simpl. gp | Comp. syll, Lgueur |
| DG | 70.00 | Phon. | non | 73.33 | Phon. | / | 100.00 | Phon. Simpl. gp | / |
| DM | 57.70* | Phon. | non | 93.33 | Phon. | / | 83.33* | Phon. | / |
| BD | 55.00* | Phon. | non | 80.00 | Phon. | Comp. syll | 33.33* | Phon. | / |

TABLE 2 : Résultats des patients aphasiques dans les tâches de dénomination, de répétition de syllabes et de mots (phon. = erreurs affectant les phonèmes , simpl. gp = simplification de groupes consonantiques, Comp. syll= effet de complexité syllabique, Lgueur = longueur). *= Performance significativement différente des participants contrôles (Crawford et al., 2010).

## 3.1 Tâches de production langagières classiques

La tâche de **dénomination d'images** (*Lexis*, Bilocq et al., 2001) consiste en une tâche de dénomination de 80 images en noir et blanc. Elle comprend des mots variant en fréquence lexicale

(faible, moyenne, élevée ; Content et al., 1990) et en longueur (1,2,3 syllabes). Les résultats sont présentés dans la Table 2. Ils montrent que les patients IJ, TM, BD, DM présentent un nombre de réponses correctes inférieur aux participants contrôles (IJ : $t(7)$= -10,041, $p$<.001; TM : $t(9)$= -2,298, $p$=.047; DM $t(9)$= -3,248, $p$=.01; BD : $t(7)$= -7,959, $p$<.001), tandis que les scores de CL et DG, quoiqu'inférieurs, ne diffèrent pas significativement des sujets sains. En outre, les patients produisent tous un grand nombre d'erreurs affectant les sons de la chaîne parlée[2] et TM présente un effet de longueur. Ensuite, nous avons proposé deux tâches **de répétition** (*Examen Long du Langage*, UCL-ULg). Dans la tâche de *répétition de syllabes*, le patient doit répéter des syllabes variant en complexité articulatoire. La tâche de *répétition de mots* consiste en une répétition de mots de différentes longueurs (1-3 syllabes) et de complexité variable. Les critères de complexité syllabique dans ces deux tâches comprennent la présence de groupes consonantiques. Les résultats et analyses qualitatives sont présentés dans la Table 2. Dans la tâche de répétition de syllabes, les résultats montrent que les patients présentent des performances dans les normes, à l'exception d'IJ, $t(7)$=-509.12, $p$<.001. Dans la tâche de répétition de mots, tous les patients, à l'exception de DG, présentent un score inférieur au groupe contrôle (IJ : $t(7)$= -471,405, $p$<.001; CL : $t(9)$= -158,942, $p$<.001; TM : $t(9)$= -211,859, $p$<.001; DM $t(9)$= -158,942, $p$<.001; BD : $t(7)$= -16,968, $p$<.001). Leurs erreurs consistent en des erreurs affectant la forme sonore des mots.

## 3.2 Tâche de répétition de non-mots destinée à analyser le VOT

Pour cette tâche, créée dans notre laboratoire, notre intérêt s'articulant autour du trait de voisement en langue française, nous avons choisi 36 non-mots CVCV, comprenant les occlusives voisées et non voisées du français /p,t,k,b,d,g/ ainsi que la voyelle /a/ (à savoir : C1V[a], C2V[a] où C1 et C2=/p,t,k,b,d,g/ ; p.ex. /gada/). Les items ont été préalablement enregistrés par une locutrice francophone avec une intonation neutre. Ils ont été présentés en ordre aléatoire au patient à l'aide d'un ordinateur PC portable à travers un casque, et il lui était demandé de les répéter. Les productions des patients ont été enregistrées à l'aide d'un enregistreur audio portable Zoom H5 avec couple stéréo en X/Y. Les contraintes de place ne nous permettant pas de détailler l'ensemble de nos analyses, nous nous centrons ici exclusivement sur la première consonne, les analyses du reste du matériel recueilli étant présentées ailleurs (Verhaegen, Delvaux, Fagniart, Huet, Piccaluga, & Harmegnies, en préparation).

En ce qui concerne le nombre et la répartition des erreurs (Figure 1), les résultats montrent que les patients aphasiques commettent tous plus d'erreurs que les sujets contrôles. De plus, alors que les erreurs les plus présentes chez les participants contrôles consistent en des absences d'explosion des plosives, des profils différents se dégagent chez les patients aphasiques. Ainsi, on remarque que les erreurs les plus présentes chez IJ et DM sont les changements de lieux d'articulation et les dévoisements, alors que TM commet autant de voisements, que de dévoisements et de changements de lieux d'articulation, que CL ne commet pratiquement que des dévoisements et DG presqu'uniquement des voisements. Enfin, on note chez BD un grand nombre de changements de lieux et de mode et de dévoisements. La figure 2 détaille les durées moyennes des VOT, en fonction des attentes voisées et non voisées des consonnes cibles chez les 6 patients ainsi que pour les groupes contrôles. Nous constatons des différences majeures entre les structures des données relatives aux patients aphasiques et celles des contrôles. D'une part, les variabilités des VOT des patients aphasiques beaucoup plus importantes que celles des sujets contrôles. Ceci suggère une grande tendance à un rapprochement des distributions de VOT des phonèmes voisés et non voisés : dans tous les cas, les intervalles de confiance à 95% se rapprochent (et même se chevauchent pour

[2] Les patients ont également tous commis des paraphasies sémantiques et présentent un effet de fréquence en dénomination d'images. Ces difficultés étant dues à des difficultés lexico-sémantiques, non centrales dans cet article, celles-ci ne seront pas discutées.

les sujets BD et IJ). On observe une tendance au raccourcissement du VOT par rapport au groupe de référence, le prévoisement tendant à être plus court dans les consonnes voisées, sauf pour le patient DG, chez qui les valeurs de VOT sont plus longues que son groupe contrôle. Les différences par rapport au groupe témoin sont significatives pour IJ, $U$= 26.50, $p$< .001, CL, $U$= 734.00, $p$= .006 et TM, $U$= 545.00, $p$= .006, mais pas pour DM, DG et BD (respectivement : $U$= 700.50, *ns* ; $U$= 981.50, *ns* ; $U$= 503.00, *ns*). Au niveau du VOT des non voisées, on note que les valeurs des VOT de TM et IJ sont proches de zéro et plus courtes que leurs contrôles (respectivement, TM : $U$= 554.00, $p$= .006 ; IJ : $U$ =346.00, p=.02) et que les valeurs deviennent même négatives chez DG, et différentes des contrôles $U$=700.50, $p$<.001. Enfin, la différenciation entre les voisées et non voisées est également nettement moins bonne chez les patients aphasiques que chez les sujets normaux. Celle-ci reste cependant significative chez tous les patients (CL : $U$= 71.00, $p$= .02; TM : $U$= 35.00, $p$= .01; DM : $U$= 36.00, $p$= .001 ; BD : $U$= 4.00, $p$= .007; DG : $U$= 85.00, $p$= .002), sauf IJ, $U$= 35.50, *ns).* Un regard plus clinique peut être porté sur le comportement des sujets au moyen de représentations telles qu'en propose la Figure 3. Nous avons déterminé, pour chaque sujet, quatre pourcentages : deux de réalisations correctes (celui des réalisations à VOT positif pour une attente de phonème non voisé et celui des réalisations à VOT négatif pour une attente de phonème voisé) et deux de réalisations incorrectes (celui des réalisations à VOT positif pour une attente de phonème voisé et celui des réalisations à VOT négatif pour un phonème non voisé). Afin de résumer cette information, nous avons calculé le coefficient d'accord de Cohen, signifiant correspondance, pour un sujet donné, entre le mode articulatoire observé et le mode articulatoire attendu. La figure de droite montre la distribution de ses valeurs pour les 6 sujets aphasiques, comparée à la distribution de l'intégralité des valeurs correspondant aux sujets de référence. On note ainsi que dans le groupe de référence, il y a peu de phénomènes de dévoisements des voisées ou de voisements des non-voisées, même si cela a tendance à augmenter chez les participants de 70-79 ans. Chez les patients aphasiques, on remarque un plus grand nombre d'erreurs, les erreurs de dévoisements constituant la tendance générale, à l'exception du patient DG qui commet plus d'erreurs de voisements que de dévoisements. Ainsi, IJ est la patiente qui commet le plus d'erreurs de dévoisements, ne présentant que 44% de voisements des voisées lorsque cela est attendu. CL présente une tendance au dévoisement également, mais de façon moins importante qu'IJ. BD, TM et DM, eux, dévoisent les voisées mais voisent également un certain nombre de non voisées.

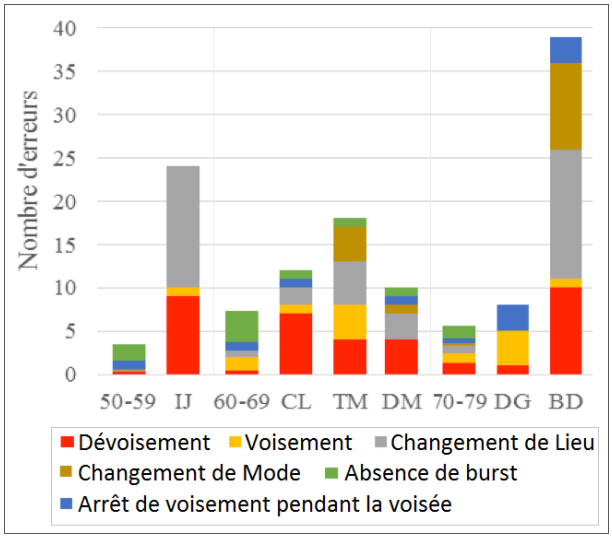



FIGURE 1. Nombre et répartition des erreurs (patients aphasiques et participants contrôle).

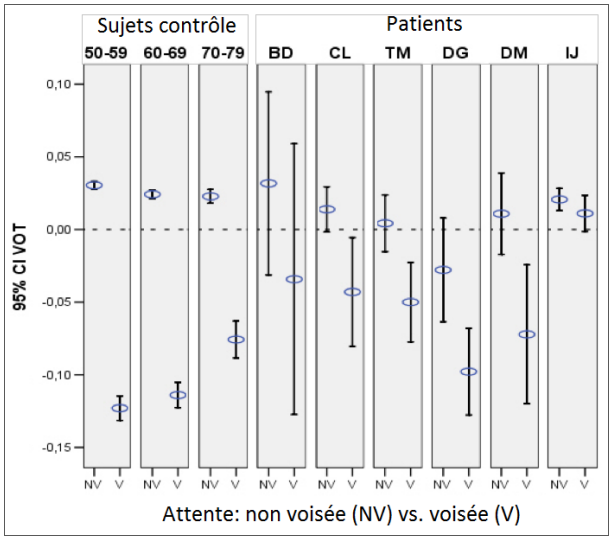



FIGURE 2 : Durées moyennes des VOT en fonction de l'attente voisée (V) ou non voisée (NV) des consonnes cibles

| Patient | Mode articulatoire Observé | Attendu non voisé | Attendu voisé | Kappa |
|---|---|---|---|---|
| BD | non voisé | 39.3% | **25.0%** | |
| | voisé | **7.1%** | 28.6% | 0.37 |
| CL | non voisé | 47.2% | **25.0%** | |
| | voisé | **2.8%** | 25.0% | 0.44 |
| TM | non voisé | 34.4% | **18.8%** | |
| | voisé | **9.4%** | 37.5% | 0.44 |
| DG | non voisé | 35.1% | **8.1%** | |
| | voisé | **16.2%** | 40.5% | 0.52 |
| DM | non voisé | 45.7% | **14.3%** | |
| | voisé | **5.7%** | 34.3% | 0.60 |
| IJ | non voisé | 47.1% | **44.1%** | |
| | voisé | **2.9%** | 5.9% | 0.06 |

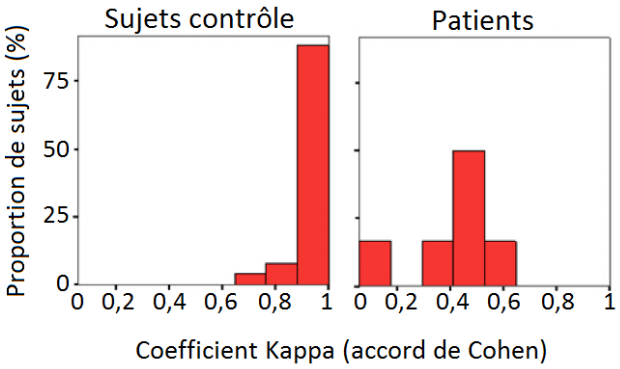



FIGURE 3 : À gauche: pourcentages des réalisations correctes et incorrectes des patients aphasiques, en fonction des attentes voisées et non voisées des consonnes cibles. À droite: distribution des coefficients kappa d'accord de Cohen entre les attentes et les réalisations.

# 4 Discussion

Par cette étude, nous souhaitions contribuer à la description des atteintes phonologiques et phonétiques dans l'aphasie. Nous avons mené une étude de cas multiples auprès de 6 patients aphasiques, quatre patients, CL, TM, IJ et DM, présentaient une aphasie non fluente et deux patients, DM et BD, présentaient une aphasie fluente. L'originalité de notre étude réside dans le fait que nous avons utilisé des techniques issues à la fois de la neuropsychologie du langage - en présentant des tâches classiques de dénomination d'images et de répétition-, mais également de la phonétique-en procédant à des analyses du Voice Onset Time des contoïdes plosives dans une tâche de répétition de non-mots. Nous avons procédé à des études de cas uniques, et non de groupes en fonction du type d'aphasie, afin de rompre avec le raisonnement circulaire, fréquent dans la littérature, qui consiste à classer les productions des patients en fonction de leurs profils prédéterminés. Les résultats montrent des profils très variables d'un patient à l'autre et parfois différents de ce que l'on pourrait rencontrer dans la littérature. IJ et CL présentent des profils qui, quoique différents sur certains aspects, tendraient vers des troubles phonétiques. Ainsi, dans les tâches de dénomination et de répétition de syllabes et de mots, les patients présentent des erreurs affectant les sons de la chaîne parlée et montrent un effet de complexité articulatoire. Dans la tâche de répétition de non-mots destinée à analyser le VOT, les profils diffèrent. En effet, les deux patients présentent un grand nombre de dévoisements de consonnes voisées (et très peu de voisements de non voisées) ainsi que des valeurs moyennes de VOT négatifs plus courtes que celles des groupes contrôles pour les voisées, traduisant des difficultés de tenue du voisement, pouvant être interprétées comme le signe de difficultés à coordonner adéquatement les articulateurs laryngés et supra-laryngés. L'analyse révèle également que le patient CL marque cependant toujours la différence entre les valeurs pour les voisées et les non voisées, même si ces valeurs sont plus variables que chez les participants contrôles. Par contre, la différence entre les voisées et non voisées chez la patiente IJ est presqu'inexistante, et les valeurs de VOT sont proches de zéro pour les deux types de plosives. L'analyse des erreurs dans cette tâche montre également un grand nombre de changements de points d'articulation chez la patiente IJ. Plus précisément, on remarque que la patiente montre une tendance forte à substituer /t/ par [k] ou [p], ce qui pourrait dès lors être dû à des difficultés d'élévation de la pointe de la langue, en raison de difficultés articulatoires. Chez les patients TM, DM et BD, l'analyse des erreurs dans la tâche de répétition de non-mots montre un grand nombre d'erreurs de changements de lieux et de modes d'articulation, ainsi que la présence d'un grand nombre de

voisements de consonnes non voisées, allant dans le sens de difficultés d'ordre phonologique chez ces patients. Ceci serait en lien avec la littérature sur les aphasies fluentes, que présentent BD et DM, mais cela est moins attendu pour le patient TM, qui est non fluent. Cependant, TM présentait également d'importantes difficultés exécutives, et l'examen clinique de son comportement lors de la tâche de répétition de non-mots nous amène à penser que ces difficultés expliqueraient une partie de ses erreurs dans cette tâche. En outre, les analyses des durées du VOT montrent des profils différents de ce qui serait classiquement attendu en cas de trouble phonologique. En effet, les trois patients présentent des valeurs de VOT beaucoup plus variables que celles des participants contrôles, la variabilité étant particulièrement importante chez BD. Leurs valeurs moyennes de VOT pour les voisées sont également plus courtes que celles des groupes de référence, ce qui pourrait traduire des difficultés de tenue de voisement et donc de coordination entre les articulateurs glottiques et supra-glottiques chez ces patients. Ceci est corroboré chez TM et BD par la présence d'un effet de complexité articulatoire en répétition de mots et de syllabes. Ainsi, les difficultés de ces patients pourraient être qualifiées de « mixtes », phonologico-phonétiques. Enfin, le patient DG présente également un profil de troubles relativement atypique. En effet, en répétition de mots et de syllabes, le patient ne présente pas d'effet de complexité articulatoire, et commet un grand nombre de voisements de consonnes non voisées. Ceci est également montré dans l'analyse des valeurs de VOT, qui montre que la majorité de ses productions tendent à être négatives et par conséquent voisées. Ces données iraient dans le sens de difficultés phonologiques. Cependant, ses valeurs de VOT restent très variables et différentes du groupe de référence, ce qui pourrait être interprété comme des difficultés d'ordre plus phonétique. L'ensemble de ces analyses montre donc la présence de profils de troubles très variables chez les patients aphasiques. Contrairement à ce qui est montré dans la littérature, les données présentées ici montrent une difficulté à classer les patients dans une catégorie de trouble, phonologique ou phonétique. En effet, la plupart des patients présentent des profils mixtes, phonologico-phonétiques. En outre, nous notons la présence de difficultés phonologiques chez des patients non fluents et de difficultés phonétiques chez des patients fluents, ce qui va à l'encontre des hypothèses généralement avancées dans la littérature (Baqué et al., 2015; Blumstein et al., 1980; Nespoulous et al., 2013; Romani et al., 2002).

Sur le plan épistémologique, la présence de ces troubles mixtes remet en question la stricte séparation entre les niveaux langagiers phonologiques et phonétiques, fréquemment présentées dans les modèles classiques de la production du langage, servant de référence notamment en clinique du langage (e.g., Levelt, 1999). Ces troubles mixtes ont également été rencontrés dans d'autres études (Galluzzi et al., 2015; Goldrick & Blumstein, 2006). Ceux-ci ont fréquemment été interprétés comme le signe de la présence de fortes interactions entre les niveaux phonologique et phonétique (Goldrick & Blumstein, 2006). Cependant, la distinction entre les niveaux phonologiques et phonétiques n'est pas présente dans tous les modèles, notamment ceux issus de la littérature en phonétique. Ainsi, une autre perspective serait d'interpréter les troubles dits « mixtes » à la lumière de théories avançant la présence d'un niveau unique phonologico-phonétique, les primitives phonologiques étant alors supposées phonétiquement spécifiées, sous la forme de gestes articulatoires à accomplir (Browman & Goldstein, 1986). Une autre piste encore pourrait amener à considérer que le niveau lexico-sémantique serait directement lié à un niveau unique de traitement des sons, où le contrôle moteur de la parole est assuré par des mécanismes complexes de feed-back et feed-forward (Hickok, Houde, & Rong, 2011). Ce type de modèles n'a cependant été encore que rarement appliqué à l'aphasie, et ils restent largement débattus.

Au niveau méthodologique, ces résultats soulignent l'intérêt d'allier des analyses acoustiques aux analyses langagières classiques, fréquemment basées sur des analyses perceptives des erreurs. En effet, certains phénomènes ne sont pas aisément perceptibles sans les analyses acoustiques. En outre, il nous semble également intéressant d'interpréter les résultats dans les tâches langagières et les analyses du VOT à la lumière d'analyses d'autres fonctions cognitives fréquemment associées au langage, telles que les fonctions exécutives (Martin, 1994), ce qui reste cependant rare. En effet, les

troubles exécutifs présentés par le patient TM nous semblent avoir fortement interféré avec ses performances dans la tâche de répétition de non-mots. Enfin, les résultats montrent également que le VOT a tendance à s'altérer avec l'âge. En effet, les résultats ont montré que le groupe de participants contrôles de 70 à 79 ans avait tendance à présenter un plus grand nombre de dévoisements de voisées que les groupes plus jeunes et que leurs valeurs moyennes de VOT étaient également plus courtes que celles des participants sains plus jeunes. Ceci amène à questionner le paramètre du VOT comme strictement indiciaire d'une pathologie langagière, étant donné que des difficultés sont également présentes dans le vieillissement sain.

En conclusion, nos résultats auprès de patients aphasiques montrent l'intérêt d'adjoindre des analyses neuropsycholinguistiques et acoustiques en vue de caractériser les troubles langagiers. Ils indiquent également la présence de profils de patients très variables et principalement « mixtes », phonologico-phonétiques, remettant en question la dichotomie classique entre les troubles phonologiques et phonétiques, fréquemment rencontrée dans la littérature en aphasiologie et qui sert également de base pour la rééducation des patients aphasiques. Ce travail montre que la distinction reste difficile à mettre en évidence dans l'analyse des patients aphasiques et invite à la réflexion sur l'utilisation d'autres modèles, tel que le modèle de la phonologie articulatoire (Browman et Goldstein, 1986), comme base d'analyse des cas des patients, en orthophonie notamment. Enfin, la présence de difficultés de voisement chez les participants âgés nous amène également à questionner le caractère strictement indiciaire d'une pathologie langagière du paramètre du VOT.

# Remerciements

Nous remercions Camille Elen, Charlotte Menu, Jérémy Pouliart, et Amélie Visentini pour leur aide dans la récolte des données.

# Références

BAQUE, L., MARCZYK, A., ROSAS, A., & ESTRADA, M. (2015). Disability, repair strategies and communicative effectiveness at the phonic level: evidence from a multiple-case study. *Neuropsycholinguistic Perspectives on Language Cognition*, (May), 144–165. https://doi.org/10.4324/9780203797365

BLUMSTEIN, S. E., COOPER, W. E., GOODGLASS, H., STATLENDER, S., & GOTTLIEB, J. (1980). Production deficits in aphasia: A voice-onset time analysis. *Brain and Language*, *9*(2), 153–170. https://doi.org/10.1016/0093-934X(80)90137-6

BROWMAN, C. P., & GOLDSTEIN, L. (1986). Towards an articulatory Phonology. *Phonology Yearbook*, *3*(1), 219–252.

CHO, T., & LADEFOGED, P. (1999). Variation and universals in VOT: evidence from 18 languages. *Journal of Phonetics*, *27*, 207–229. https://doi.org/10.1006/jpho.1999.0094

DELL, G. S., SCHWARTZ, M. F., MARTIN, N., SAFFRAN, E. M., & GAGNON, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*(4), 801–38. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9337631

GALLUZZI, C., BURECA, I., GUARIGLIA, C., & ROMANI, C. (2015). Phonological simplifications, apraxia of speech and the interaction between phonological and phonetic processing. *Neuropsychologia*, *71*, 64–83. https://doi.org/10.1016/j.neuropsychologia.2015.03.007

GOLDRICK, M., & BLUMSTEIN, S. E. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes*, *21*(1), 649–683. https://doi.org/10.1080/01690960500181332

HICKOK, G., HOUDE, J., & RONG, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron*, *69*(3), 407–22. https://doi.org/10.1016/j.neuron.2011.01.019

INDEFREY, P. (2011). The spatial and temporal signatures of word production components: a critical update.

*Frontiers in Psychology*, *2*, 255. https://doi.org/10.3389/fpsyg.2011.00255
KUROWSKI, K., & BLUMSTEIN, S. E. (2016). Phonetic basis of phonemic paraphasias in aphasia: Evidence for cascading activation. *Cortex*, *75*, 193–203. https://doi.org/10.1016/j.cortex.2015.12.005
LEVELT, W. J. (1999). Models of word production. *Trends in Cognitive Sciences*, *3*(6), 223–232. https://doi.org/10.1016/S1364-6613(99)01319-4
LISKER, L., & ABRAMSON, A S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word Journal Of The International Linguistic Association*.
NESPOULOUS, J. L., BAQUE, L., ROSAS, A., MARCZYK, A., & ESTRADA, M. (2013). Aphasia, phonological and phonetic voicing within the consonantal system: preservation of phonological oppositions and compensatory strategies. *Language Sciences*, *39*(1), 117–125.
PILLON, A., & DE PARTZ, M.-P. (2014). Sémiologie, syndromes aphasiques et examen clinique des aphasies. In X. Seron & M. Van der Linden (Eds.), *Traité de neuropsychologie clinique de l'adulte. Tome 1 -Evaluation (2ème édition)* (De Boeck-S, pp. 249–265). Paris.
RAPP, B., & GOLDRICK, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review*, *107*(3), 460–99.
ROMANI, C., OLSON, A., SEMENZA, C., & GRANA, A. (2002). Patterns of phonological errors as a function of a phonological versus an articulatory locus of impairment. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *38*(4), 541–67.
RYALLS, J., PROVOST, H., & ARSENAULT, N. (1995). Voice onset time production in French-speaking aphasics. *Journal of Communication Disorders*, *28*(1), 205–215.
VERHAEGEN, C., DELVAUX, V., FAGNIART, S., HUET, K., PICCALUGA, M., & HARMEGNIES, B. (en préparation). Phonetic and/or phonological paraphasias in aphasia: an acoustic study of speech output in six aphasic patients.

# Prénasalisation des plosives voisées initiales dans la parole apraxique : une étude aérodynamique pilote

Anna Marczyk[1, 2], Yohann Meynadier[1], Maria-Josep Solé[3]
(1) Aix Marseille Université, CNRS, LPL, Aix-en-Provence, France
(2) Brain and Language Research Institute, Aix-en-Provence, France
(3) Universitat Autònoma de Barcelona, Barcelona, Espagne
anna.marczyk@lpl-aix.fr, yohann.meynadier@lpl-aix.fr,
mariajosep.sole@uab.cat

## RESUME

Cette étude exploratoire examine le débit d'air nasal dans la production de consonnes plosives chez une locutrice apraxique hispanophone, dans le but de déterminer si elle utilise la nasalité comme mécanisme compensatoire pour faciliter l'initiation du voisement. Les résultats mettent en évidence que les plosives initiales identifiées comme [+voisé] présentent un débit d'air nasal significativement plus élevé que les consonnes [-voisé] chez cette locutrice, ce qui va dans le sens de l'hypothèse d'une prénasalisation comme un mécanisme adaptatif face au déficit du voisement typique de cette pathologie. Cette stratégie est particulièrement favorisée (i) en position initiale de phrase et (ii) par la présence d'une consonne nasale suivant la plosive cible, reposant sur une anticipation de nasalité de la coda dans l'attaque syllabique (ex. ***b**anco*). Par ailleurs, l'absence d'effet de la fréquence lexicale sur l'occurrence de la prénasalisation des plosives [+voisé] suggère que cette compensation opèrerait au niveau phonétique d'encodage de la parole.

## ABSTRACT

**Prenasalization of word-initial stops in apraxia of speech: a preliminary aerodynamic study.**
This preliminary study examines nasal leak during the production of Spanish word-initial stops in an apraxic speaker in order to determine if this motor adjustment is used as a compensatory mechanism aimed at facilitating the initiation of voicing in stops. The results show that /b d g / correctly identified as 'voiced' exhibit significantly larger amounts of nasal airflow during the stop closure than those identified as 'voiceless' ([-voice] and devoiced stops). These results suggest that prenasalization may be an adaptive mechanism to overcome the frequent devoicing errors of apraxia. Nasal leak during voiced /b d g/ –but not their voiceless counterparts– is significantly larger (i) in phrase initial vs non-phrase initial position, and (ii) when followed by tautosyllabic nasal (e.g. ***b**anco*). On the other hand, no effect of word frequency on the occurrence of nasal leak in voiced stops was found suggesting that this compensatory strategy takes place at the phonetic level of encoding.

MOTS-CLES : apraxie de la parole, débit d'air nasal, prénasalisation, voisement, compensation.
KEYWORDS: apraxia of speech, nasal airflow, prenasalization, voicing, compensation.


10.21437/JEP.2018-77

# 1 Introduction

L'apraxie verbale est un trouble de la parole d'origine neurologique qui affecte sélectivement les processus d'encodage des gestes articulatoires (Buckingham et Christman, 2008 ; Ogar, Slama, Dronkers, Amici, et Gorno-Tempini, 2005 ; Ziegler, 2002 *inter alia*). En particulier, les travaux phonétiques sur différentes langues ont mis en lumière un déficit du contrôle laryngé et de la synchronisation entre articulateurs glottique et supraglottiques (Auzou et al. 2000, Blumstein et al. 1980, Verhaegen et al. 2016), provoquant des fréquentes erreurs d'assourdissement des obstruantes [+voisé], spécifiquement en position initiale. Néanmoins, due aux contraintes aérodynamiques spécifiques au contexte initial (après pause), la mise en vibration des plis vocaux est également critique pour des sujets sains qui montrent eux aussi des stratégies d'ajustement moteur favorisant le voisement des obstruantes en initiale. En espagnol notamment, on observe de fréquentes prénasalisations spécifiques aux consonnes plosives [+voisé] en début de phrase (Solé, 2009 ; Solé, 2018). Cette manœuvre permettant d'évacuer la pression intra-orale favoriserait ainsi le flux d'air transglottique nécessaire à l'initiation de la vibration laryngée.

Une étude acoustique de la parole de deux locutrices hispanophones présentant une apraxie verbale pure a permis d'observer des durées anormalement longues du pré-voisement des plosives [+voisé] (i.e. un VOT négatif) et des oscillations de grande intensité, pouvant être attribuées à la réalisation d'un échappement nasal compensatoire (Marczyk et al. 2017). Les analyses acoustiques effectuées jusqu'alors ne permettent pas d'identifier de manière non équivoque le recours à un ajustement articulatoire nasal. L'objectif de cette étude pilote est de valider l'existence réelle de cet échappement nasal pour l'une des deux patients apraxiques de notre étude acoustique à partir de l'enregistrement des débits d'air. Les hypothèses générales suivantes sont questionnées :

H1. *Hypothèse de l'absence de déficit moteur*. Les sujets apraxiques ne présentent pas de déficit strictement moteur qui affecte la fonction vélo-pharyngienne. Notre hypothèse prédit que puisqu'ils contrôlent bien les gestes vélo-pharyngiens, les émissions nasales seront différentes pour /m/, /b/, /p/.

H2. *Hypothèse gestuelle*. L'émission nasale relève de l'anticipation ou de la persévération de l'abaissement du voile du palais associé à la consonne nasale qui suit ou précède. Cette hypothèse prédit que les plosives, [+voisé] comme [-voisé], sont nasalisées seulement et toujours quand une consonne nasale est dans leur voisinage.

H3. *Hypothèse cognitive*. Les processus de sélection phonologique seraient préservés dans l'apraxie : en position initiale du mot les sujets apraxiques peuvent tenter d'établir le contraste phonologique de voisement au moyen de la nasalisation. Si cette hypothèse est infirmée, on observera une émission nasale que pour les plosives [+voisé], et non pour les [-voisé].

H4. *Hypothèse de facilitation contextuelle.* Si l'hypothèse cognitive (H3) est corroborée, deux contextes susceptibles de faciliter la prénasalisation des plosives [+voisé] seront investigués : (i) la coarticulation progressive et (ii) la coarticulation régressive.

# 2 Méthodologie

## 2.1 Corpus

Les mots contenant les plosives cibles ont été insérés en phrase porteuse (Ha dicho *el* ***b****ote* dos veces, 'Il a dit *le pot* deux fois'). Le locuteur a été instruit de lire uniquement le(s) mot(s) cible, présentés en caractère gras, sur l'écran d'ordinateur. Chaque mot a été produit 5 fois par chaque locuteur (donnant lieu à la variable *Item*).

Le corpus comprend une liste de 200 stimuli manipulant :

(i) la catégorie phonémique : plosive [-voisé] (**t**aza ‘tasse’) vs. plosive [+voisé] (**d**ato ‘donnée’) vs. consonne nasale (**n**ada ‘rien’) en position lexicale initiale

(ii) le lieu d’articulation : labial (**p**alo ‘bâton’, **b**ote ‘pot’, **móvil** ‘téléphone portable’) vs. coronal (**d**ato ‘donnée’, **t**aza ‘tasse’, **n**ada ‘rien’) vs. dorsal (**g**ato ‘chat’, **c**asa ‘maison’)

(iii) la fréquence lexicale : mots fréquents (**g**ato ‘chat’) vs. mots peu fréquents (**g**asa ‘gaz’) vs. mots sans signification (**g**apo)

(iv) le contexte phonétique : position initiale post-pausale (**d**ato, **t**aza, **n**ada) vs. position intervocalique (to**d**o ‘tout’, pa**t**o ‘canard’, ca**n**a ‘cheveu blanc’)

(v) la coarticulation nasale progressive en frontière de syllabe (et de mot) où /n/ précède une plosive [+voisé] (u<u>n</u> **b**ote ‘un pot’) ou [-voisé] (u<u>n</u> **p**ote ‘une marmite’) vs. où /l/ précède une plosive [+voisé] (e<u>l</u> **b**ote ‘le pot’) ou [-voisé] (e<u>l</u> **p**ote ‘la marmite)

(vi) la coarticulation régressive en interne de syllabe (et de mot): la plosive [+voisé] (**b**anco ‘banque’, **d**onde ‘où’, **g**ancho ‘crochet’) ou [-voisé] (**p**ongo ‘je mets’, **t**onto ‘bête’, **c**anto ‘chant’) précède /n/ en position codaïque homosyllabique vs. la plosive [+voisé] (**b**ote ‘bidon’, **d**ato ‘donnée’, **g**ato ‘chat’) ou [-voisé] (**p**ote ‘pot’, **t**aza ‘tasse’, **c**asa ‘maison’) n’est pas suivie d’une consonne nasale en coda homosyllabique.

## 2.2 Enregistrement et acquisition aérodynamique

Une locutrice apraxique hispanophone (34 ans, droitière) a participé à cette étude pilote. Le diagnostic de l’apraxie de la parole a été établi par une orthophoniste à l’hôpital de Bellvitge (Barcelone, Espagne), et il a été confirmé par une évaluation postérieure effectuée au moment de la passation de l’épreuve. Ses troubles langagiers sont survenus suite à la résection d'une tumeur de type gliome impliquant l’opercule frontal dans l'hémisphère gauche. L'évaluation clinique n'a révélé aucun déficit de dénomination, de compréhension ou de répétition, le trouble apraxique étant l’unique séquelle postopératoire. Ce trouble se caractérise par un débit de parole ralenti et de fréquentes erreurs d’assourdissement des plosives sonores, particulièrement en position initiale. Une femme sans aucune pathologie du langage, appariée en âge (38 ans) et en niveau éducatif a été également enregistrée pour les mêmes tâches afin de servir de sujet contrôle.

Les données ont été acquises avec la station de travail EVA2 développée au Laboratoire Parole et Langage (Ghio & Teston 2002), constituée d’un micro-ordinateur, auquel sont connectés des capteurs acoustiques et aérodynamiques. Cette instrumentation non invasive permet d’acquérir synchroniquement au signal acoustique les débits d’air nasal et oral calibrés. L’acquisition se fait via l’interface d’acquisition multi-canaux de Phonedit[1], qui sert aussi de logiciel d’étiquetage et d’analyse des signaux. Chaque participant a été enregistré en deux sessions d’environ 20 mn séparées par une pause de 10 mn. En outre, la pression intra-orale a été mesurée en glissant une sonde entre les lèvres du sujet.

## 2.3 Segmentation et mesures

La segmentation et l’étiquetage phonologique ont été effectués manuellement à partir du signal acoustique synchronisé avec les données aérodynamiques (Figure 1). Pour les consonnes sonores (plosives [+voisé] et nasales) initiales après pause, le début de l’occlusion a été déterminé à l’aide de la représentation oscillographique (notamment à partir de l’amplitude correspondant au début des

[1] http://www.lpl-aix.fr/~lpldev/phonedit.

pulsations glottiques). Pour déterminer le début de l'occlusion des plosives [-voisé] et [+voisé] dévoisées, nous nous sommes basés sur la dynamique de la courbe correspondant au débit oral (DAB). Ainsi, une courbe du débit d'air buccal descendante et une augmentation continue de la pression orale (pour les bilabiales) sont indicatives d'une fermeture de la bouche pendant l'occlusion. Sur les portions du signal ainsi segmentées, nous avons pris la mesure du débit d'air nasal moyen durant la production de la plosive cible.

Il convient de noter que selon cette méthode de segmentation, les phénomènes ayant lieu aux frontières des segments entrent dans le calcul. Ainsi par exemple, les plosives se trouvant au voisinage d'une consonne nasale peuvent présenter des valeurs positives du débit nasal dû à la coarticulation ou tout simplement au temps nécessaire pour la fermeture du voile du palais (cf. le début de [k] de *banco*, Figure 1). Dans le cas des plosives [-voisé] ou [+voisé] dévoisées, un DAN positif peut être interprété comme une nasalisation résiduelle n'affectant pas la perception [-voisé] de ces consonnes.

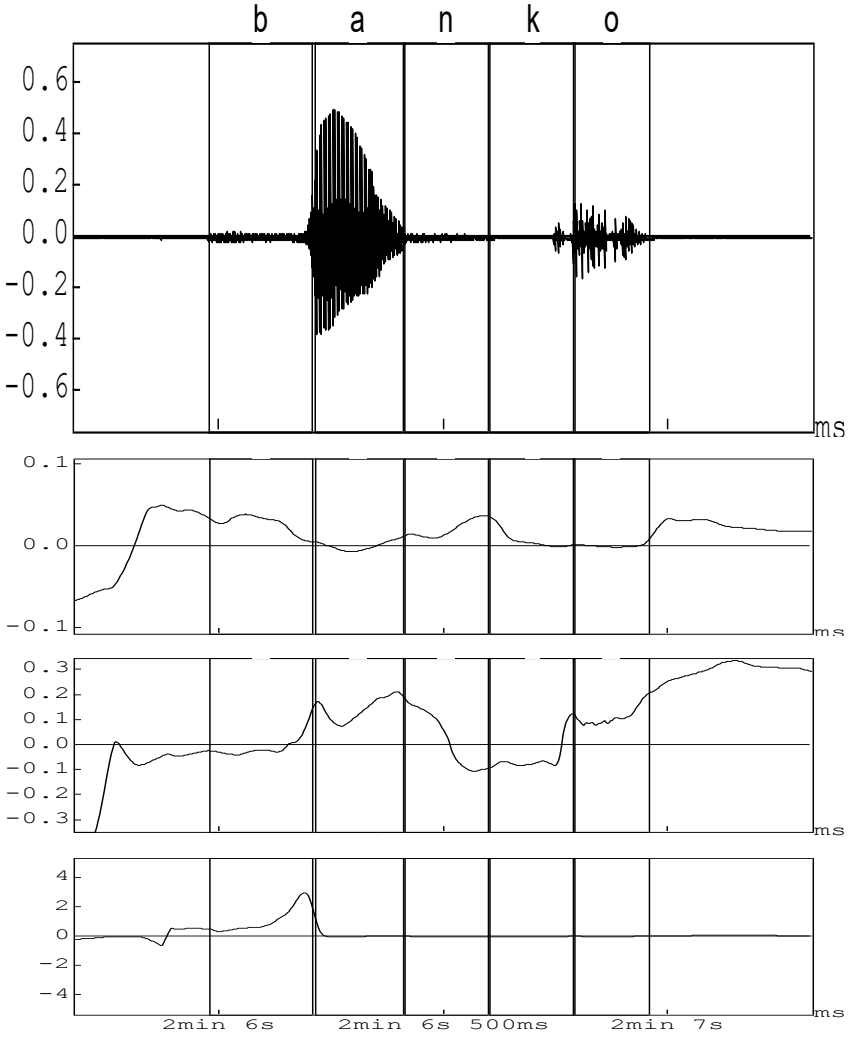



FIGURE 1 : Segmentation du mot *banco* 'banque', en position post-pausale, produit par la locutrice apraxique. De haut en bas, l'onde acoustique, le DAN (en $dm^3/s$), le DAB (en $dm^3/s$) et la pression intra-orale (en hP).

### 2.4 Prétraitement et analyses statistiques

Tous les mots ont été transcrits (orthographiquement) par deux juges indépendants. De plus, les plosives /b d g/ ont été perceptivement classées comme 'sonores' si elles ont été perçues [+voisé] et comme 'dévoisée' si perçues [-voisée] (e.g. *cava* ou ['kaβa] est perçu comme ['kapa]).

L'effet des variables indépendantes sur les valeurs de débit d'air nasal (ci-après DAN) a été statistiquement analysé par des modèles mixtes de régression linéaire avec le DAN comme variable réponse et items (mot*répétitions) comme facteur aléatoire. Dans chaque modèle, nous avons introduit un à un les prédicteurs de l'étude, à savoir : le *Locuteur* (apraxique vs. contrôle), la *Catégorie phonétique* (plosive sonore vs. dévoisée vs. sourde vs. consonne nasale), le *Contexte phonétique (##CV, Vl#CV, ##CVn, Vn#CV, VCV*) et la *Fréquence lexicale* (cf. §2.1). Les valeurs moyennes de débit obtenues via EVA ont été converties en ml/s.

## 3 Résultats

Parmi 105 obstruantes [+voisé] produites par la locutrice apraxique, 75 (71%) ont été identifiées comme dévoisées (erreur). Ces productions se caractérisaient toutes par l'absence de périodicité. Parmi les plosives [+voisé] dévoisées, 22 concernent des bilabiales (29%), 17 les dentales (23%) et 36 des vélaires (48%). Donc seules 29% des plosives [+voisé] ont été correctement perçues, à la différence de toutes (100%) des plosives sourdes et nasales.

Les résultats relatifs au facteurs *Locuteur* et *Catégorie phonétique* sont présentés dans la Figure 2. Ils montrent d'intéressantes asymétries entre la locutrice apraxique et contrôle. La première concerne les erreurs d'assourdissement, présentes chez la locutrice avec apraxie et absentes chez la locutrice contrôle. La seconde est relative aux valeurs du débit d'air nasal : celles-ci sont supérieures pour les consonnes nasales pour toutes les deux locutrices mais, de manière intéressante, les valeurs de ce paramètre sont plus importantes pour les plosives sonores que pour les sourdes chez la locutrice apraxique et pas chez la locutrice contrôle. Ce dernier résultat suggère l'existence d'un mécanisme « nasal » facilitant l'initiation du voisement chez la locutrice apraxique.

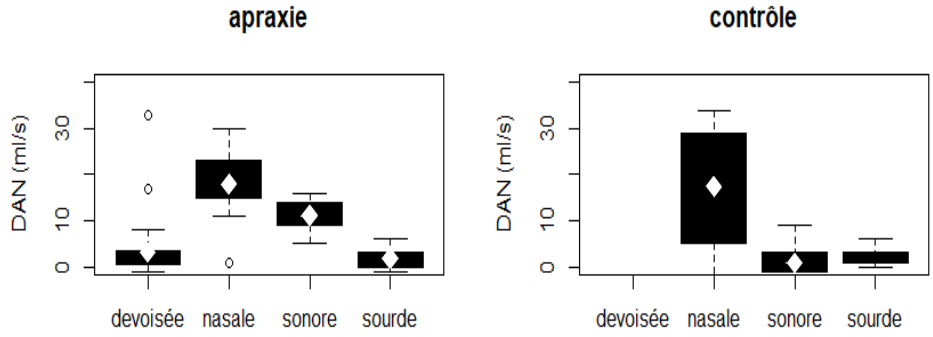



Figure 2 : Distribution des valeurs du débit d'air nasal en fonction de la *Catégorie phonétique* en position initiale du mot chez las locutrice apraxique et contrôle. Locutrice apraxique : N = 200 (75 dévoisées, 20 nasales, 30 sonores, 75 sourdes). Locutrice contrôle : N=200 (0 dévoisée, 20 nasales, 105 sonores, 75 sourdes).

Ces observations sont confirmées par les analyses statistiques. La *Catégorie phonétique* montre une interaction significative avec le facteur *Locuteur* ($F_{(2,128)}$=7.92, p=.000), toutes fréquences lexicales confondues. Les analyses séparées pour la locutrice apraxique montrent un effet significatif de la *Catégorie phonétique* ($F_{(3,62)}$=25.25, p=.000) et révèlent des différences significatives entre les consonnes nasales et (i) les plosives sonores (18 ml/s (1.27) vs. 11 ml/s (1.18), $t_{(62)}$=2.72, p=.041),

(ii) les plosives dévoisées (3.11 ml/s (1.12), $t_{(62)}$=7.66, p=.000) et (iii) les plosives sourdes (1.86 ml/s (1.54), $t_{(62)}$=7.28, p=.001).

Surtout, les analyses montrent un DAN moyen significativement plus important pour les plosives sonores que pour les dévoisées ($t_{(63)}$=3.97, p=.001) et les sourdes ($t_{(62)}$=4.06, p=.001). Par contre, les plosives dévoisées ne se différencient pas significativement des plosives sourdes ($t_{(62)}$=0.75, p=.877). Cela indique que seules les plosives initiales sonores sont prénasalisées chez le sujet apraxique, et que la prénasalisation n'est pas produite pour les consonnes dont la spécification phonologique est [-voisé].

La locutrice saine (contrôle) ne présente aucun cas de dévoisement de plosives sonores en initiale de mot. Mais surtout, la prénasalisation des plosives sonores est aussi faible que celle relevée pour les plosives sourdes (absence de différence significative entre ces deux catégories, $t_{(62)}$=0.86, p=.663). Cela indique que le sujet contrôle n'a pas recours à la prénasalisation pour assurer la vibration laryngée des plosives sonores en initiale de mot.

La *Fréquence lexicale* du mot (cf. § 2.1, iii) ne montre aucun effet significatif sur le DAN des plosives et nasales initiales chez aucune des locutrices ($t_{(2,128)}$=0.21, p=.806), indiquant que la prénasalisation est indépendante de la lexicalité et de la familiarité des mots produits.

La Figure 3 représente la distribution selon le *Contexte phonétique* (et syllabique, cf. §2.1, iv-vi) des valeurs du DAN pour les plosives sonores (correctement perçues (et produites) [+voisé], en haut) et pour les plosives perçues [-voisé] (incluant les [-voisé] et les [+voisé] dévoisées perçues comme [-voisé], en bas). Comme noté plus haut (Figure 2), les consonnes sourdes et dévoisées ne présentant pas de différences pour le DAN, elles ont été donc regroupées. L'effet de la présence d'une consonne nasale au voisinage de la plosive cible est ici testé s'agissant de la prénasalisation des plosives.

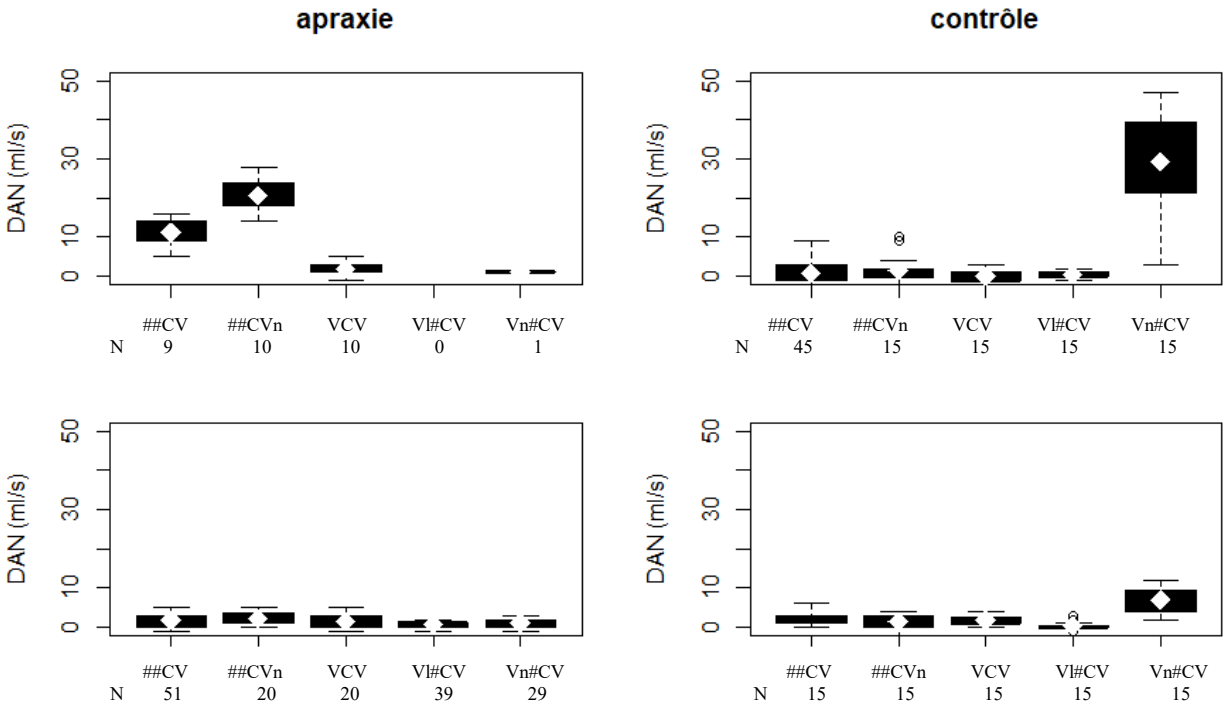



FIGURE 3 : Distribution des valeurs du DAN pour les plosives sonores (en haut) et les plosives perçues [-voisé] (en bas) selon le contexte phonétique. Les effectifs (N) sont donnés sous chaque contexte pour les locutrices apraxique (à gauche) et contrôle (à droite).

Les différences entre les deux locutrices sont confirmées par un effet de l'interaction entre *Contexte phonétique* et *Locuteur* pour les plosives sonores ($F_{(3,126)}$=27.44, p=.000, Figure 3 en haut) et pour les plosives perçues [-voisé] ($F_{(4,210)}$=9.45, p=.000, Figure 3 en bas). Les analyses du DAN séparées par locuteur montrent un effet simple significatif du *Contexte phonétique* pour les plosives sonores tant chez la locutrice apraxique que chez la locutrice saine (respectivement, $F_{(3,23)}$=53.16, p=.000 et $F_{(4,100)}$=84.22, p=.000). Pour les plosives perçues [-voisé], un effet significatif n'est présent que chez la locutrice contrôle ($F_{(4,65)}$=25.44, p=.000). Mais surtout les contextes de prénasalisation des plosives sont très différents selon les locutrices, indiquant des stratégies spécifiques.

La locutrice saine présente un DAN significativement plus élevé exclusivement dans le contexte où la plosive, [+voisé] comme [-voisé], est précédée par une consonne nasale (Vn #CV, un **b**ote). Dans ce contexte, le DAN moyen est significativement plus élevé pour les plosives sonores que pour les plosives sourdes ($F_{(1,28)}$=40.59, p=.000). Aucune différence significative n'apparaît entre tous les autres contextes. Il ressort donc que la prénasalisation chez la locutrice contrôle n'est pas liée à l'initiation du voisement en position initial (##CV(n)), mais est issue d'un processus de nasalisation progressive. Par ailleurs, cette coarticulation de nasalité est en bonne partie associée au voisement, comme le signale un DAN bien plus élevé pour les plosives sonores que pour les sourdes. Autrement dit, l'élévation du voile survient plus tard pendant l'articulation de la plosive sonore (comparé à la plosive sourde) ou bien, les différences de valeurs sont dues à différences d'amplitude entre ces deux contextes. Ce mécanisme contribue à maintenir un différentiel de pression sus- et sous-glottique favorable à la vibration laryngée pendant l'occlusion de la plosive.

Au contraire, pour la locutrice apraxique la prénasalisation semble être spécifiquement produite pour soutenir l'initiation du voisement. Le DAN est significativement plus élevé pour les plosives sonores en position initiale (#CV) que dans tous les autres contextes : vs. VCV, $t_{(22)}$=5.81, p=.001 ; vs. Vn #CV, $t_{(25)}$=2.81, p=.044. Egalement, dans cette position, la présence d'une consonne nasale en coda de syllabe (##CVn, **b**a<u>n</u>co), favorise un DAN significativement plus élevé par rapport au contexte sans consonne nasale codaïque (##CV, **b**ote) : $t_{(22)}$=6.09, p=.000. De plus, à la différence de la locutrice contrôle, la présence d'une consonne nasale avant la plosive (Vn #CV, u<u>n</u> **b**ote) ne favorise pas un DAN significativement plus élevé. La locutrice apraxique ne produit pas d'assimilation nasale progressive des plosives, qu'elles soient [+voisée] ou [-voisé]. L'ensemble de ces résultats, à savoir (i) le DAN plus élevé au contexte avec une consonne nasale codaïque par rapport à ce même contexte sans coda et (ii) des valeurs non significatives du DAN de la plosive sonore si la consonne nasale qui la précède appartient à une syllabe différente, suggère que la syllabe joue un rôle important chez les locuteurs apraxiques.

# 4 Discussion et conclusion

Nos résultats sont congruents avec les analyses acoustiques de Marczyk et al. (2017), confirmant la présence des plosives voisées prénasalisées dans des contextes phonétiques spécifiques chez une locutrice apraxique.

Considérant maintenant nos hypothèses, ces résultats cadrent parfaitement avec l'hypothèse de l'absence d'un déficit strictement moteur qui pourrait affecter le contrôle du port vélo-pharyngien chez la locutrice apraxique (H1). En effet, la prénasalisation ne concerne que les plosives [+voisé], et non les [-voisé], dans une position spécifique, à savoir en initiale de mot. En outre, les consonnes nasales sont, elles, bien produites et perçues. L'hypothèse gestuelle (H2), développée dans le cadre de la Phonologie Articulatoire, soutenant que l'émission nasale durant la plosive (sourde comme sonore) est essentiellement mécanique, n'est pas corroborée par notre étude. En effet, le voisinage d'une nasale ne permet pas seul d'expliquer l'intrusion d'un geste d'abaissement vélaire indépendamment de la spécification du trait de voisement de la plosive. La Figure 3 (en haut)

montre ainsi que seules les plosives sonores (à savoir spécifiées et perçues comme [+voisé]) en position initiale, précédant ou non une consonne nasale homosyllabique (##CV(n)), sont produites avec un débit d'air nasal. Au contraire, les plosives perçues comme [-voisé] ne sont jamais nasalisées dans les mêmes contextes (Figure 3, en bas). Néanmoins, le fait qu'en contexte pré-nasal (#CVn) un débit d'air nasal est plus important qu'en contexte non nasal (##CV) est en accord avec l'hypothèse de facilitation contextuelle de la prénasalisation pour les plosives sonores initiales de mot (H4). Il apparaitrait ainsi que la planification gestuelle puisse jouer un certain rôle dans le processus de compensation du voisement par nasalisation chez les locuteurs apraxiques.

Enfin, l'hypothèse cognitive (H3) rend assez bien compte de nos résultats. En effet, l'abaissement du voile du palais semble être contrôlé par des spécifications phonologiques segmentales et positionnelles. Seule la position initiale de mot déclenche la prénasalisation des seules plosives [+voisé] chez notre sujet apraxique. Ce processus pourrait donc bien être une manœuvre de compensation planifiée du voisement phonologique, indiquant qu'une sélection cognitive des phonèmes et des contextes ciblés opère. La finalité de la prénasalisation serait alors de permettre l'initiation de la vibration laryngée dans une positon prosodique défavorable. Néanmoins, le pourcentage élevé de plosives [+voisé] dévoisées (71% des cas), à savoir sans maintien du voisement par prénasalisation, suggère tout de même l'existence d'une forte composante motrice dans le trouble apraxique. Dans ces cas, on peut penser que le sujet n'a pas pu compenser le dévoisement des plosives [+voisé] par un mécanisme de nasalisation.

Par ailleurs, notre étude n'est qu'en partie compatible avec les résultats de Solé (2009, 2018) sur la prénasalisation des plosives [+voisé] chez les locuteurs sains hispanophones. Solé (2018) met en lumière une série de mécanismes moteurs observés pour la production du voisement des plosives en position initiale de phrase, à savoir un échappement nasal ou oral, ou un élargissement du conduit vocal. Or, la locutrice contrôle de notre étude ne produit aucune plosive [+voisé] prénasalisée dans cette position. Reste qu'on ne peut exclure que notre locutrice n'ait recours aux autres ajustements moteurs décrits par Solé (2018).

Pour finir, les résultats de notre étude diffèrent notablement de ceux rapportés par des études antérieures quant à la coarticulation de voisement chez les patients apraxiques (Katz, 1988 ; Ziegler et von Cramon, 1985). Ces études rapportent un déficit de coarticulation régressive chez ces patients. Or, le fait que la locutrice apraxique de notre étude puisse pré-nasaliser les plosives initiales de mot quand une consonne nasale codaïque homosyllabique est présente (par ex. **b**a<u>n</u>.co vs. **b**o.te), peut faire penser que la coarticulation régressive serait préservée chez ces patients. Ce résultat est congruent avec l'hypothèse d'un rôle important joué par la syllabe en tant qu'unité d'encodage dans la production verbale de la locutrice apraxique. Il est compatible avec les résultats indicatifs de l'accès préservé aux programmes moteurs des syllabes (Aichert & Ziegler, 2004). Il correspond par ailleurs à la perception de la parole apraxique, qui se caractérise par un débit d'élocution lent et une tendance à la scansion syllabique (Edmonds & Marquardt, 2004). Pour avancer sur ces questions soulevées par la présente étude, la nécessité d'investigations plus larges semble une nouvelle étape fondamentale pour des recherches sur la parole et la phonologie des sujets apraxiques.

# Remerciements

Nous remercions Ana María Fernández Planas, Université de Barcelone. Etude financée par ANR-11-LABX-0036 (BLRI), ANR-11-IDEX-0001-02 (A*MIDEX), FFI2017-84479-P du Ministère de la Science et de l'Innovation espagnol.

# Références


AICHERT, I. AND ZIEGLER, W. The role of the syllable in apraxia of speech: Theoretical background, empirical observations and therapeutic consequences. *Forum Logopadie* 2004. Vol. 18(2):6-13.

AUZOU P., ÖZSANCAK C., MORRIS R. J., JAN M., EUSTACHE F., HANNEQUIN D. « Voice onset time in aphasia, apraxia of speech and dysarthria: a review ». *Clinical Linguistics & Phonetics.* 2000. Vol. 14, n°2, p. 131-150.

BLUMSTEIN S. E., COOPER W. E., GOODGLASS H., STATLENDER S., GOTTLIEB J. « Production deficits in aphasia: A voice-onset time analysis ». *Brain Lang.* 1980. Vol. 9, n°2, p. 153–170.

BUCKINGHAM H. W., CHRISTMAN S. « Disorders of Phonetics and Phonology ». In : STEMMER B, WHITAKER HA, ÉD. *Handbook of the Neuroscience of Language* London : Academic Press Elsevier, 2008. p. 127-136.

EDMONDS L. & MARQUARDT T. «Syllable use in apraxia of speech: Preliminary findings». *Aphasiology*. 2004. 18:12, 1121-1134.

GHIO A., TESTON B. « Caractéristiques de la dynamique d'un pneumotachographe pour l'étude de la production de la parole: aspects acoustique et aérodynamique ». In : *24èmes Journées d'Etudes sur la Parol.* [s.l.] : [s.n.], 2002.

KATZ W. F. « Anticipatory coarticulation in aphasia: Acoustic and perceptual data ». *Brain and Language* 1988. Vol. 35, n°2, p. 340–368.

KUROWSKI K., BLUMSTEIN S. E., PALUMBO C. L., WALDSTEIN R., BURTON M. « Nasal Consonant Production in Broca's and Wernicke's Aphasics: Speech Deficits and Neuroanatomical Correlates ». *Brain and Language* 2008. Vol. 100, n°3, p. 262–275.

MARCZYK A., MEYNADIER Y., GAYDINA Y., SOLÉ M.-J. « Dynamic acoustic evidence of nasalization as compensatory mechanism for voicing in Spanish apraxic speech ». In : *Proc. 17 ISSP Tianjin, China*. [s.l.] : Springer, 2017.

OGAR J., SLAMA H., DRONKERS N., AMICI S., GORNO-TEMPINI M. L. « Apraxia of speech: an overview ». *Neurocase*. 2005. Vol. 11, n°6, p. 427-432.

SOLÉ M.-J. « Acoustic and aerodynamic factors in the interaction of features. The case of nasality and voicing ». In : VIGÁRIO M, FROTA S, FREITAS MJ, ÉD. *Phonetics and Phonology: Interactions and Interrelations* [s.l.] : [s.n.], 2009. p. 205–234.

SOLÉ M.-J. « Articulatory adjustments in initial voiced stops in Spanish, French and English. » *Journal of Phonetics* 2018. Vol. 66, p. 217-241.

VERHAEGEN C., DELVAUX V., HUET K., FAGNIART S., PICCALUGA M., HARMEGNIES B. « La distinction entre les paraphasies phonétiques et phonologiques dans l'aphasie: Etude de cas de deux patients aphasiques ». In : *Actes des Journées d'Etudes sur la Parol.* Paris : [s.n.], 2016.

ZIEGLER W. « Psycholinguistic and motor theories of apraxia of speech ». *Semin. Speech Lang.* [En ligne]. 2002. Vol. 23, p. 231-243. Disponible sur : < http://dx.doi.org/10.1055/s-2002-35798 >

ZIEGLER W., VON CRAMON D. « Anticipatory coarticulation in a patient with apraxia od speech ». *Brain and Language.* 1985. Vol. 26, p. 117-130.

# La « voyelle apicale » en chinois de Jixi : caractéristiques acoustiques et comportement phonologique

Bowei Shao & Rachid Ridouane
Laboratoire de Phonétique et Phonologie, (CNRS / Sorbonne Nouvelle),
19 Rue des Bernardins, 75005 Paris, France
bowei.shao@univ-paris3.fr, rachid.ridouane@univ-paris3.fr

## RESUME

Cette étude s'intéresse aux propriétés phonétiques et phonologiques de la « voyelle apicale » /z/, telle qu'elle est attestée en chinois de Jixi. L'objectif est de déterminer si ce segment a des propriétés d'une voyelle ou d'une consonne. Phonologiquement, ce segment est un phonème distinct qui s'appose à /i/. Il est exclusivement attesté en position noyau de syllabe où il constitue une unité porteuse de ton, et peut subir des processus de sandhi tonal. Phonétiquement, nous avons examiné les caractéristiques acoustiques de ce segment en se basant sur la production de 10 locuteurs natifs. Les résultats obtenus montrent que /z/ contient dans la majorité des cas du bruit de friction dans sa phase initiale, mais une structure formantique apparaissant vers sa fin. L'analyse du rapport bruit/harmonique confirme cette présence importante du bruit, distinguant ainsi clairement ce segment des voyelles hautes /i, u, ʉ/. Ces résultats soulèvent un ensemble de questions sur la relation entre la réalisation phonétique d'un segment et son comportement phonologique.

## ABSTRACT

**"Apical vowel" of Jixi-Hui Chinese: acoustic characteristics and phonological behavior.**
This study investigates the phonetic and phonological properties of the "apical vowel" /z/, as attested in Jixi-Hui Chinese. Our goal is to determine if this segment has the properties of a vowel or a consonant. Phonologically, this segment is a distinct phoneme from /i/. It is exclusively attested in syllable nucleus position where it constitutes a tone-bearing unit, and can undergo tonal sandhi processes. Phonetically, we examined the acoustic characteristics of this segment based on the production of 10 native speakers. The results obtained show that /z/ contains in the majority of cases friction noise in its initial phase, but a formant structure appears towards its end. The analysis of the harmonic-to-noise ratio confirms this significant presence of noise, clearly distinguishing this segment from the high vowels /i, u, ʉ/. These results raise a series of questions about the relationship between the phonetic realization of a segment and its phonological behavior.

MOTS-CLES : Voyelles apicales, langue chinoise de Jixi, phonétique, phonologie.
KEYWORDS: Apical vowels, Jixi-Hui Chinese, phonetics, phonology.

# 1 Introduction

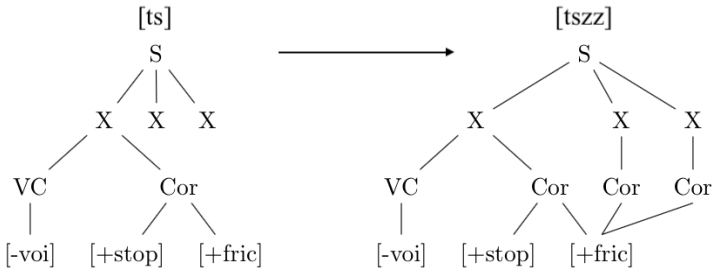
## 1.1 Les « voyelles apicales » du chinois standard : bref aperçu

Les langues chinoises sont connues pour avoir une série de segments spécifiques, sous le nom de « voyelles apicales ». En chinois standard (CS), ils apparaissent dans des syllabes qui ont pour attaques des sibilantes coronales et sont homorganiques avec celles-ci.


10.21437/JEP.2018-78

Ces « voyelles apicales » du CS, qui peuvent être dentales ou rétroflexes, ont fait l'objet de plusieurs études. L'appellation « voyelles apicales » est proposée pour la première fois par Karlgren (1915). Selon d'autres auteurs, il s'agirait plutôt de « voyelles fricatives » (Ladefoged, Maddieson, 1996) ou d'« approximantes syllabiques » (Lee, Zee, 2003 ; Lee-Kim, 2014). Pour Yu (1999), ces segments sont des « sibilantes syllabiques » et l'indication de leur propriété de sibilante est la présence de bruit dans les hautes fréquences (4000 Hz – 6000 Hz). Cependant, Lee-Kim (2014) montre qu'il n'y a aucun bruit de friction dans les hautes fréquences chez la majorité des locuteurs qu'elle avait enregistrés. A l'évidence, il n'existe pas encore de consensus sur la caractérisation phonétique de ces segments, tantôt considérés comme des voyelles et tantôt comme des consonnes.

Au niveau phonologique, les « voyelles apicales » du CS sont analysées comme des allophones en distribution complémentaire avec la voyelle [i] : elles apparaissent uniquement suivant les sibilantes coronales [s, ts, $ts^h$, ʂ, tʂ, $tʂ^h$], alors que la voyelle /i/ peut être précédée des sibilantes palatales [ɕ, tɕ, $tɕ^h$] et d'autres consonnes. Dell (1994) propose d'analyser ces segments comme des « fricatives syllabiques » ou plus exactement comme « le prolongement voisé » de leurs attaques sibilantes. Duanmu (2007) propose une analyse similaire en considérant la « voyelle apicale », non pas comme un segment indépendant, mais comme résultant de la propagation du trait [+ fricative] vers un noyau vide (Duanmu 2007 : 44) :




Traditionnellement, les « voyelles apicales » sont transcrites avec les symboles [ɿ] pour la dentale et [ʅ] pour la rétroflexe. Outre ces symboles, qui ne sont pas adoptés par l'API, les études mentionnées ci-dessus proposent d'autres symboles de transcription, dont voici un tableau récapitulatif :

| | | |
|---|---|---|
| ɿ ʅ | voyelles apicales | (Karlgren, 1915 ; Zee, Lee, 2007) |
| z̩ z̩ | sibilantes syllabiques | (Yu, 1999) |
| ɹ̩ | approximante syllabique | (Lee, Zee, 2003) |
| z z̩ | fricatives syllabiques | (Dell, 1994 ; Duanmu, 2007) |
| ɹ̪̩<br>ɻ̩ | approximante dentale syllabique<br>approximante rétroflexe syllabique | (Lee-Kim, 2014) |

TABLE 1 : Les transcriptions proposées pour les « voyelles apicales » du chinois standard.

Dans notre étude sur la langue chinoise de Jixi, une seule « voyelle apicale » est concernée et nous adoptons la notation [z], suivant en cela Dell (1994) et Duanmu (2007).

## 1.2 La langue chinoise de Jixi et sa « voyelle apicale »

La langue chinoise de Jixi 绩溪 (CJ) est une langue du groupe huizhou 徽州, parlée dans le district Jixi, au sud de la province de l'Anhui 安徽 (Zhao, 1989, 2003 ; Hirata, 1998). Les descriptions qui existent sur cette langue sont faites à partir de la variante parlée dans la ville de Jixi. Notre étude est aussi basée sur des données émanant de cette variante.

Le CJ a huit voyelles monophtongues [i, y, u, ʉ, o, ɤ, ɔ, a] et une « voyelle apicale » [z], issue diachroniquement de la voyelle /i/ selon Zhu (2004). Contrairement au CS, /z/ et /i/ en CJ sont aujourd'hui deux phonèmes distincts, comme le montrent les paires minimales suivantes :

[tsz$^{31}$] ‘poule’ *vs*. [tsi$^{31}$] ‘un nom de famille’ ; [ts$^{h}$z$^{31}$] ‘déception’ *vs*. [ts$^{h}$i$^{31}$] ‘automne’ ; [sz$^{31}$] ‘soie’ *vs*. [si$^{31}$] ‘réparer’.

De plus, /z/ a une distribution plus large, il apparaît dans les syllabes [pz, p$^{h}$z, mz, nz, tsz, ts$^{h}$z, sz] et ces syllabes sont lexicalement développées, comptant pour 7,2% des entrées monosyllabiques du dictionnaire du CJ (Zhao, 2003). Comme le montrent ces syllabes, /z/ n’est pas systématiquement homorganique avec les consonnes attaques. Les bilabiales [p, p$^{h}$, m] et la nasale dentale [n], n’ayant pas de caractéristique sibilante à prolonger, suggèrent que /z/ ne peut être analysé comme une prolongation voisée de l’attaque, mais comme un phonème à part.

Etant noyau de syllabe, /z/ est une unité porteuse de ton, qui peut subir des processus phonologiques de sandhi tonal, comme le montre l’exemple en (1).

(1) /**sz$^{31}$**/ ‘ouest’ + /ko$^{31}$/ ‘melon’ → [**sz$^{33}$**ko$^{31}$] ‘pastèque’

Dans les proverbes rimés et les poèmes, les syllabes ayant /z/ comme noyau ne riment qu’avec d’autres syllabes qui ont le même noyau (voir 2, où les tons sont ignorés).

(2) /mɔ.tɛ̃.mɔ.**ts$^{h}$z**/
‘les champs se vendent’

/mɔ.pɤʔ.tɔ.si.**nz**/
‘les savoir faire ne se vendent pas’

Très peu de travaux ont été menés sur la « voyelle apicale » en CJ. Pour Zhao (1989) l’inventaire phonémique du CJ contient une consonne /z/ qui ne peut être suivie que de la « voyelle apicale » comme noyau. Dans une description plus récente, Zhao (2003) a supprimé la consonne /z/ de l’inventaire phonémique de la langue, en soulignant que « la voyelle apicale » commence toujours avec un léger bruit de friction. Une autre description (Hirata, 1998) mentionne brièvement le cas du /z/ du CJ, en indiquant que quand la « voyelle apicale » se trouve dans une syllabe sans attaque, elle se réalise avec un bruit de friction, ce qui fait selon les auteurs qu’elle ressemble à une consonne fricative [z].

### 1.3 Problématique

Les études précédentes sur la « voyelle apicale » [z] sont faites essentiellement sur le CS, où [z] est une variante contextuelle de la voyelle /i/. Ce n’est pas le cas en CJ. Comme nous l’avons montré plus haut, la « voyelle apicale » /z/ en CJ se comporte phonologiquement comme une voyelle : elle s’oppose lexicalement à /i/, peut occuper le noyau de syllabe et porte les tons. Cependant, les dialectologues remarquent aussi que ce segment peut contenir du bruit de friction, une caractéristique de consonne.

Dans la suite de ce travail nous cherchons à déterminer les caractéristiques phonétiques de ce segment. Nous cherchons plus spécifiquement à déterminer si ces segments ont des caractéristiques acoustiques d’une voyelle ou d’une consonne fricative. Notre objectif est d’essayer d’expliquer la relation entre la réalisation phonétique de /z/ en CJ et sa spécification phonologique.

## 2 Méthodologie

Les données acoustiques examinées dans cette étude émanent de dix locuteurs natifs, cinq hommes (M1 – M5) et cinq femmes (W1 – W5). Ils sont nés entre 1964 – 1974 (âge moyen : 49 ± 3,8). Tous les locuteurs ont grandi dans la ville de Jixi, avec leurs parents qui sont eux aussi nés et ont grandi dans la même ville. Ils parlent la même variante du CJ dans leurs milieux familiaux et

professionnels et se considèrent comme locuteurs natifs sans accent.

Nous avons établi une liste de monosyllabes avec /a, i, u, ʉ, z/ comme noyaux, et avec /p, $p^h$, m, n, ts, $ts^h$, s/ comme attaques. Ces syllabes ont des tons différents et forment des mots réels, inclus dans la phrase cadre en (3) :

(3) /$ki^{44}$ ɛɔ$^{213}$ _ ɛɔ$^{213}$ sɔ$^{44}$ fa$^{44}$/
'Il écrit _ trois fois'

Les sessions d'enregistrement sont effectuées dans la ville de Jixi, dans un studio de télévision locale, à l'aide d'un micro-casque hypercardioïde (AKG C520), une carte son Edirol UA25 et le logiciel Audacity, version 2.1.0. La liste de mots est répétée entièrement cinq fois par chaque locuteur, donnant lieu au total à 2150 syllabes cibles ([z] : 550, [i] : 400, [u] : 450, [a] : 350, [ʉ] : 400). Les syllabes cibles sont segmentées manuellement en utilisant Praat (Boersma, Weenink, 2017). Etant donnée la difficulté de déterminer clairement la frontière entre attaque et noyau pour les syllabes /tsz, $ts^h$z, sz/, nous avons pris la première « *pulse* » détectée par Praat comme début de la partie voisée du /z/ (voir FIGURE 1 à droite).

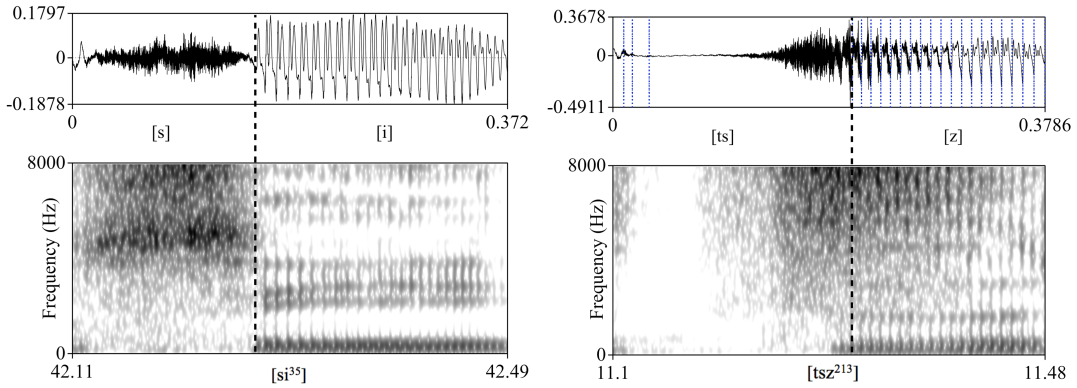



FIGURE 1 : Exemples de segmentation, [$si^{35}$] (à gauche) prononcé par M4 et [$tsz^{213}$] prononcé par W3. Les lignes fines pointillées sur le signal acoustique à droite représentent les « *pulses* » détectées par Praat.

Un ensemble de caractéristiques a été examiné en partant des signaux acoustiques et des spectrogrammes.

1. Nous avons inspecté de manière visuelle la présence et l'absence du bruit de friction dans les segments /z/ ;
2. Nous avons mesuré les formants des voyelles /a, i, u, ʉ/ et du /z/ sur le deuxième tiers de chaque segment concerné ; le formant maximum (F5) est à 5000 Hz pour les hommes et 5500Hz pour les femmes, avec une fenêtre temporelle (Gaussian) de 0,025s.
3. Nous avons analysé le rapport bruit/harmonique pour /i, ʉ/ et pour /z/, quantifié grâce au Harmonics-to-Noise Ratio (HNR).

Les formants sont calculés sur la partie centrale des segments cibles et le HNR a été calculé sur la durée totale de ces segments.

# 3 Résultats

## 3.1 Présence et absence de bruit de friction

L'examen des signaux acoustiques et des spectrogrammes montre que 34% des segments /z/

contiennent du bruit de friction sur plus de la moitié de leurs durées ; 53% moins de la moitié et 13% ont un bruit de friction quasi-invisible. La nature de la consonne précédente, qu'elle soit labiale ou coronale, orale ou nasale, ne semble pas avoir d'effet important sur la quantité de bruit de friction observée.

Les syllabes /pz/ et /nz/ illustrent les cas typiques de la réalisation du /z/ avec du bruit de friction sur plus de la moitié de la durée (voir FIGURE 2). Dans la syllabe /pz/, il est peu probable que le bruit de friction soit une prolongation de l'attaque /p/, qui est une occlusive bilabiale. De manière similaire, dans la syllabe /nz/ (FIGURE 2 à droite), l'attaque /n/ ne contient normalement aucun bruit de friction. Malgré la possibilité d'analyser les bruits de friction comme une prolongation de l'attaque pour les syllabes /tsz, ts$^{h}$z, sz/, la présence de bruit de friction dans les syllabes /pz, p$^{h}$z, mz, nz/ ne peut pas être expliquée par la même raison. Les bruits de friction des segments /z/ doivent donc être considérés, au moins pour certaines syllabes, comme des caractéristiques intrinsèques au segment /z/, et non pas comme une prolongation de l'attaque, comme cela a été proposé pour le CS.

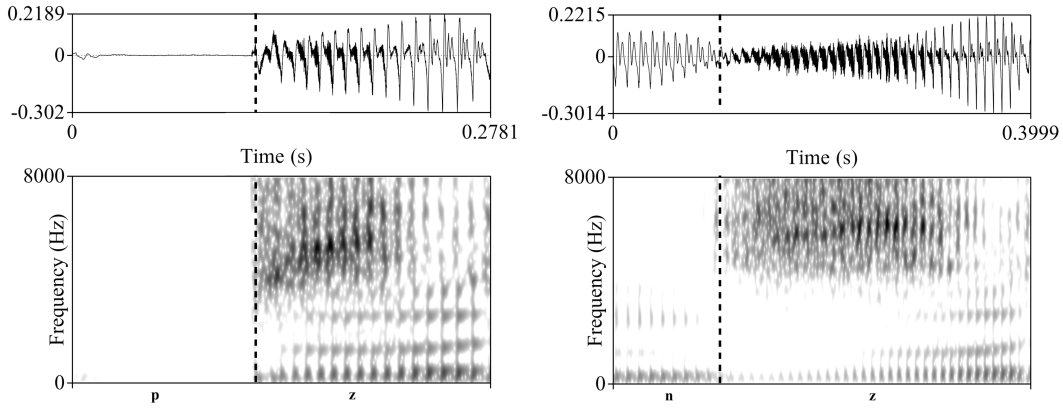



FIGURE 2 : Exemples de /z/ contenant du bruit de friction avec attaque /p/ (à gauche) et /n/, les deux prononcés par M3.

Nous pouvons aussi constater que le bruit de friction ne se prolonge jamais sur l'ensemble de la durée de /z/ ; la structure formantique devenant de plus en plus nette à mesure que le bruit de friction diminue vers la fin du segment. En attendant des analyses articulatoires par échographie programmées pour le futur proche, nous proposons d'expliquer cette phase de bruit suivie de formants comme suit : pendant la phase initiale de la tenue de /z/, la pointe de la langue s'approche du palais, formant un passage d'air étroit qui donne lieu à une forte turbulence. La pointe de la langue baisse par la suite, le passage d'air s'élargit, le bruit de friction diminue progressivement, et les formants apparaissent.

## 3.2 Sur la structure formantique de /z/

Nous avons montré en 3.1 que le segment /z/ peut contenir du bruit de friction sur une grande partie de sa durée, ressemblant en cela à une consonne fricative. Cependant, ce segment se comporte phonologiquement comme une voyelle, comme montré en 1.2. Dans cette partie, nous proposons d'analyser la structure formantique de /z/. Les valeurs de F1 et F2 des segments /a, i, u, ʉ, z/ sont montrés en FIGURE 3 et les moyennes sont données en TABLE 2. Nous constatons que /z/ a un F1 bas et un F2 central, des caractéristiques d'une voyelle centrale fermée. Cependant, nous constatons qu'il existe un chevauchement important entre l'ellipse du /z/ et l'ellipse du /ʉ/.

Le test ANOVA sur l'ensemble des données (femmes et hommes ensemble) indique que les valeurs de F1 pour /z, ʉ, u/ sont significativement différentes ($F$(2, 1249)=199,6 ; $p$<0,001), le test post-hoc TukeyHSD montre que c'est dû à la différence entre /z/ et /ʉ/ ($p$<0,001), et à la différence entre /u/ et /ʉ/ (p<0,001), alors qu'il n'y a pas de différence significative entre /z/ et /u/ ($p$>0,5). Le test post-hoc Student-Newman-Keuls montre le pattern suivant : F1(u)=341 Hz, F1(z)=338 Hz > F1(ʉ)=284 Hz.

Pour les valeurs de F2, le test ANOVA indique qu'il existe une différence significative entre /z, ʉ, u/ ($F(2, 1249)=386,9$ ; $p<0,001$). Le test post-hoc TukeyHSD montre que les trois segments /z, ʉ, u/ ont des F2 différents ($p=0$). Le test post-hoc Student-Newman-Keuls montre le pattern suivant : F2(z)=1366 Hz > F2(ʉ)=1222 Hz > F2(u)= 844 Hz.

| | Voyelles et /z/ du CJ | | | | | | | | | | [z] du CS | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Femmes | | | | | Hommes | | | | | Femmes | Hommes |
| | /a/ | /i/ | /z/ | /ʉ/ | /u/ | /a/ | /i/ | /z/ | /ʉ/ | /u/ | /z/ | /z/ |
| F3 | 2721 | 3444 | 3217 | 2904 | 2901 | 2423 | 3097 | 2608 | 2309 | 2408 | 3500,84 | 2922,58 |
| F2 | 1589 | 2810 | 1643 | 1336 | 932 | 1195 | 2149 | 1094 | 1108 | 748 | 1680,32 | 1295,94 |
| F1 | 972 | 320 | 358 | 311 | 366 | 755 | 262 | 319 | 256 | 313 | 376,28 | 396,50 |

TABLE 2 : Valeurs moyennes (Hz) de F1, F2, F3 de / a, i, u, ʉ, z/ en CJ, comparées avec les valeurs du /z/ en CS d'après Zee et Lee (2001).

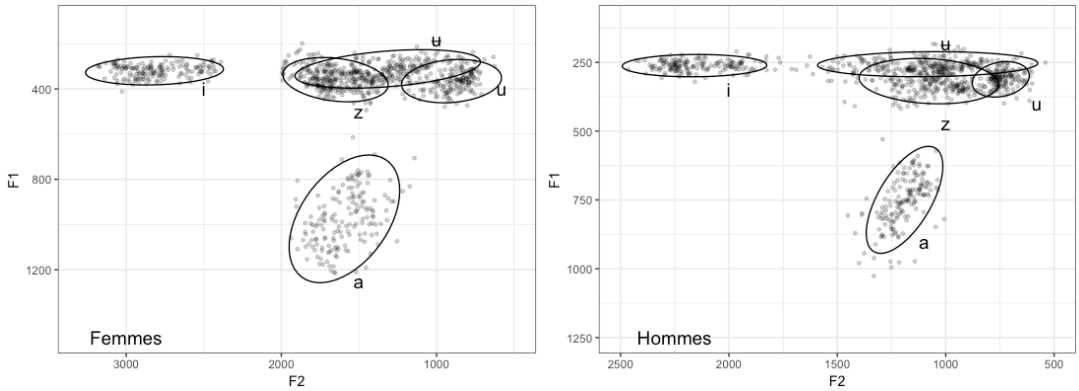



FIGURE 3 : Ellipses vocaliques de /a, i, u, ʉ, z/ pour femmes (à gauche) et hommes.

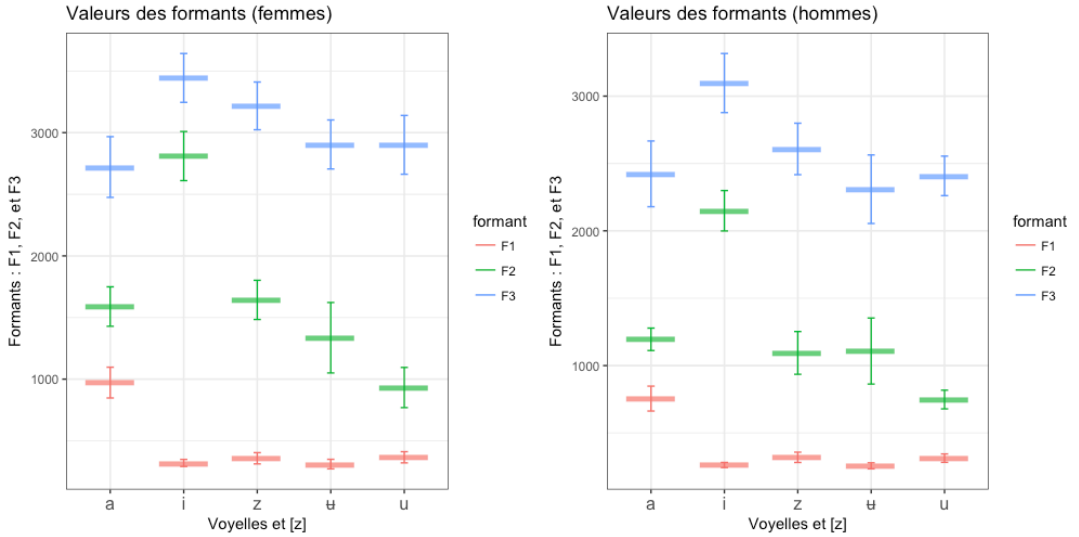



FIGURE 4 : Valeurs moyennes de F1, F2, F3 de /a, i, u, ʉ, z/ pour femmes (à gauche) et hommes. Les barres d'erreur représentent l'écart-type.

Concernant les valeurs de F3, le test ANOVA indique qu'il existe une différence significative entre /z, ʉ, u/ ($F(2,1249=93,39$ ; $p<0,001$). Le test post-hoc TukeyHSD indique que c'est dû à la différence entre /z/ et /ʉ/ ($p=0$), et à la différence entre /z/ et /u/ ($p=0$), alors qu'il n'y a pas de

différence significative entre /u/ et /ʉ/ ($p$>0,05). Le test post-hoc Student-Newman-Keuls montre le pattern suivant : F3(z)=2910 Hz > F3(u)=2666 Hz, F3(ʉ)= 2609 Hz.

Les mesures des formants (TABLE 2 et FIGURES 3, 4) et les tests statistiques montrent que la voyelle /ʉ/ est caractérisée par un F1 très bas, la voyelle /u/ est caractérisée par un F2 très bas, le segment /z/ a un F1 bas, un F2 moyen et un F3 très élevé. Comparé au [z] du CS (Zee, Lee, 2001), le /z/ du CJ a plus ou moins les mêmes caractéristiques formantiques. Il est intéressant de signaler, par ailleurs, que ces valeurs formantiques de /z/ correspondent aussi à celles obtenues pour la fricative [z] du polonais et de l'anglais américain (Jassem, 1965)[1]. Une telle ressemblance peut suggérer que /z/ du CJ, au moins d'un point de vue acoustique, peut aussi être traité comme une consonne fricative [z].

### 3.3 Rapport bruit/harmonique (HNR)

Comme nous l'avons vu, la majorité des réalisations de /z/ en CJ contient du bruit de friction. Nous avons mesuré le HNR de /z/ et nous l'avons comparé à celui de /i, ʉ/ (i 5). Nous avons choisi /i/ pour pouvoir comparer nos résultats avec ceux obtenus par Faytak (2015), qui a mesuré le HNR de /z, i/ du CS. Nous avons aussi choisi d'inclure /ʉ/ dans nos analyses car /z/ et /ʉ/ présentent un chevauchement important dans l'espace vocalique (voir FIGURE 3 ci-dessus).

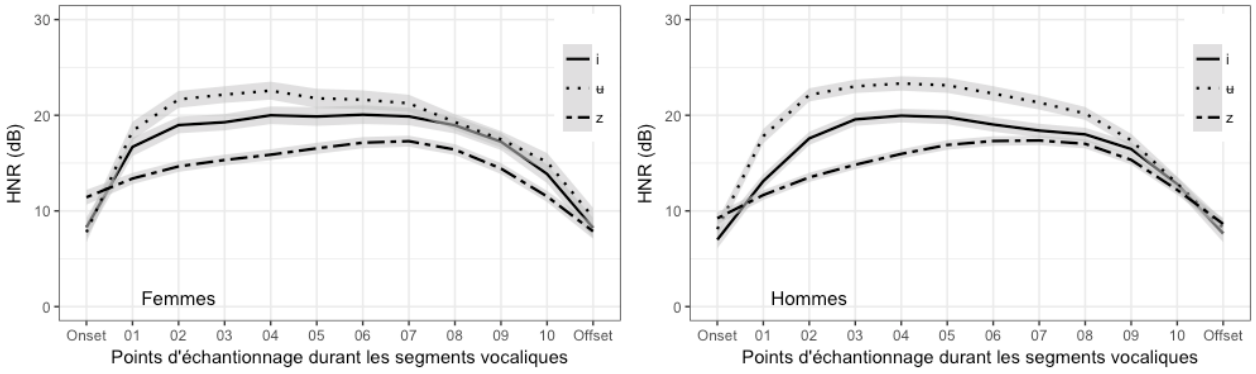



FIGURE 5 : HNR de /i, ʉ, z/ pour femmes (à gauche) et hommes. Pour chaque segment, les lignes centrales représentent la moyenne et les zones grises représentent l'intervalle de confiance de 95%.

Les voyelles /i, ʉ/ ont des valeurs HNR qui montent vite au début de la voyelle, les valeurs restent stables et forment une sorte de plateau entre point 2 et point 7, ces valeurs descendent vers la fin des voyelles. Les valeurs HNR de /z/ indiquent un pattern différent. Au début du segment, les valeurs HNR commencent à un niveau légèrement plus élevé que pour /i, ʉ/, ensuite la montée est beaucoup plus lente, sans aucun plateau entre le début du segment et le sommet, ensuite la phase de descente ressemble à celles de /i, ʉ/. Ce résultat fournit un argument supplémentaire indiquant que /z/ contient plus de bruit de friction que /i, ʉ/. Il est aussi compatible avec la description articulatoire que nous avons proposée pour rendre compte de ce bruit de friction suivie d'une structure formantique pour /z/ : la pointe de la langue s'abaisse progressivement durant la tenue de /z/ et le bruit de friction cesse, correspondant à la montée progressive des valeurs d'HNR.

## 4 Discussion et conclusion

Cette étude a permis de dégager un ensemble de caractéristiques phonologiques et phonétiques du segment /z/ du CJ. Phonologiquement, il se comporte comme une voyelle : il s'oppose lexicalement

[1] Polonais : F2 1770 Hz, F3 2870 Hz ; anglais américain : F2 1570 Hz, F3 2720 Hz. Les mesures ont été effectuées sur un locuteur pour chaque langue, le F1 n'a pas été mesuré et le sexe des locuteurs n'a pas été indiqué.

aux autres voyelles (et notamment à /i/) ; il joue le rôle de noyau de syllabe ; il est une unité porteuse de ton et peut subir des processus de sandhi tonal. Phonétiquement, il présente des caractéristiques hybrides : il contient très souvent une forte quantité de bruit de friction que nos valeurs HNR confirment. En cela il ressemble à une consonne fricative. Mais ce segment affiche aussi une structure formantique nette apparaissant vers sa fin, affichant ainsi des caractéristiques typiques d'une voyelle.

La comparaison de ce segment à la voyelle apicale en CS montre des ressemblances et des différences : Du point de vue niveau phonétique, les deux segments ont une structure formantique assez semblable, mais la présence du bruit de friction en CS n'a pas été constatée par tous les chercheurs, suggérant que le /z/ en CJ a probablement plus de bruit de friction qu'en CS. Du point de vue phonologique, le /z/ en CJ, contrairement au CS, ne peut être analysé comme une variante contextuelle de /i/. De même, contrairement au CS, il ne peut non plus être analysé comme un prolongement voisé de l'attaque, puisqu'il est attesté aussi bien suivant sibilantes dentales que suivant consonnes labiales et/ou nasales, des consonnes qui n'ont pas de caractéristique sibilante qu'elles peuvent objectivement propager.

La forte présence du bruit de friction pendant la tenue de /z/ en CJ suggère qu'il s'agit d'une consonne fricative voisée [z]. La présence de ce bruit de friction dans la majorité des cas est probablement une stratégie utilisée par les locuteurs pour maximiser la distance perceptive entre /z/ et /ʉ/. Nous avons en effet relevé que le segment /z/ a une structure formantique qui ressemble à celle de la voyelle /ʉ/, or cette ressemblance peut potentiellement créer une confusion perceptive, et doit donc être évitée. Phonologiquement, le segment /z/ se trouve uniquement en position noyau de syllabe, favorisant en cela une analyse de ce segment comme « voyelle apicale ». La présence d'une structure formantique à la fin de ce segment est probablement développée pour accomplir la fonction d'unité porteuse de ton. Néanmoins, la possibilité d'analyser ce /z/ comme une consonne ne peut pas pour autant être exclue. D'une part parce que la présence des formants n'est pas une propriété exclusive des voyelles : les consonnes sonantes, entre autres, ont aussi des formants. De même la fricative [z] des autres langues non apparentées peuvent avoir une structure formantique qui ressemble à celle observée pour le /z/ du CJ. D'autre part, parce qu'en CJ, d'autres consonnes peuvent aussi être noyaux de syllabes et des unités porteuses de ton (*e.g.* [m, n, v]). Pour autant, une différence importante doit être signalée : [m, n, v] se trouvent non seulement en position de noyau de syllabe, mais aussi en position attaque, contrairement à /z/ qui ne peut être que noyau.

L'examen des caractéristiques phonétiques et phonologiques de ce segment doit à l'évidence être poursuivi. Les différents arguments développés ici semblent suggérer que l'analyse de /z/ comme une consonne rend compte de ces propriétés phonétiques (notamment la forte présence du bruit de friction pendant sa tenue). Une telle analyse peut aussi être accommodée à son comportement phonologique : cette consonne serait semblable à d'autres consonnes comme /v/, par exemple, qui peut aussi fonctionner comme noyau de syllabe. Pour autant, la façon dont ce segment fonctionne phonologiquement doit encore être examinée de plus près. Pourquoi la distribution de ce segment est limitée à la position noyau ? A-t-on affaire à une consonne fricative ou à une consonne approximante ? De même ses propriétés articulatoires doivent d'être examinées. Des analyses par échographie sont ainsi prévues pour examiner la configuration linguale de ce segment : est-elle semblable à celle des voyelles, des fricatives ou des approximantes ? Ces questions et d'autres sont intéressantes et peuvent nous renseigner sur la relation entre forme phonétique d'un son et sa fonction phonologique, et plus globalement sur l'organisation et la malléabilité des systèmes phonétiques/phonologiques.

# Remerciements

Nous tenons à remercier tous les locuteurs qui ont participé à l'acquisition des données. Cette recherche s'insère dans le programme « Investissements d'Avenir » géré par l'Agence Nationale de la Recherche ANR-10-LABX-0083 (Labex EFL).

# Références


BOERSMA P., WEENINK D. (2017). Praat: doing phonetics by computer (version 6.0.36) [Logiciel]. Repéré à http://www.praat.org/.

DELL, F. (1994). Consonnes à prolongement syllabique en Chine. *Cahiers de linguistique-Asie orientale* 23(1), 87-94.

DUANMU S. (2007). *The phonology of standard Chinese*. New York : Oxford University Press.

FAYTAK M. (2015). Temporal organization of frication in fricativized vowels. Poster présenté au *169th ASA Meeting*, 18-22 mai 2015, Pittsburgh PA.

HIRATA S. (1998). *Huizhou Fangyan Yanjiu* [Etude sur les dialectes du Huizhou]. Tokyo : Kohbun Press.

JASSEM, W. (1965). The formants of fricative consonants. *Language and Speech* 8(1), 1-16.

KARLGREN B. (1915). *Etudes sur la phonologie chinoise*. Uppsala : KW Appelberg.

LADEFOGED P., MADDIESON I. (1996). *The sounds of the world's languages*. Oxford & Malden, MA : Blackwell.

LEE W. S., ZEE E. (2003). Standard Chinese (Beijing). *Journal of the International Phonetic Association* 33(1), 109-112.

LEE-KIM S. -I. (2014). Revisiting Mandarin 'apical vowels': An articulatory and acoustic study. *Journal of the International Phonetic Association* 44(3), 261-282.

YU A. (1999). Aerodynamic constraints on sound change: The case of syllabic sibilants. *The Journal of the Acoustical Society of America* 105(2), 1096-1097.

ZHAO R. (1989). Anhui Jixi Fangyan Yinxi Tedian [Caractéristiques phonologiques du dialecte jixi de province Anhui]. *Fangyan* 2, 125-130.

ZHAO R. (2003). *Jixi Fangyan Cidian* [Dictionnaire du dialecte jixi]. Nanjing : Jiangsu Jiaoyu Chubanshe.

ZEE, E., & LEE, W. S. (2001). An acoustical analysis of the vowels in Beijing Mandarin. Actes d'*INTERSPEECH*, 643-646.

ZEE, E., & LEE, W. S. (2007). Vowel typology in Chinese. Actes de *16th International Congress of Phonetic Sciences*, 1429-1432.

ZHU, X. (2004). Hanyu yuanyin de gao ding chu wei [Changements diachroniques des voyelles fermées dans les dialectes chinois]. *Zhongguo Yuwen*, 302(5), 440-451.

Delhoume Anaïs
Delvaux Véronique
Demolin Didier
Desnous Florent
Didirkovà Ivana
Dodane Christelle
Dohen Marion
Dos Santos Christophe
Doukhan David
Duchemin Angéline
Dufour Richard
Dufour Sophie
Estève Yannick
Fagniart Sophie
Farinas Jérôme
Fauth Camille
Feng Gang
Ferragne Emmanuel
Ferré Gaëlle
Ferreira Sébastien
Fougeron Cécile
Fredouille Corinne
Gaillard Pascal
Garnier Maëva
Gendrot Cédric
Gerber Silvain
Ghannay Sahar
Ghio Alain
Giusti Laurence
Grabli David
Grenez Francis
Gresse Adrien
Gros-Bonfiglioli Audrey
Guiraud Hélène
Guitard-Ivent Fanny
Hallé Pierre
Harmegnies Bernard
Herment Sophie
Herry-Bénit Nadine
Hincapié Ana-Sofia
Hirsch Fabrice
Huang Yizhi
Huet Kathy
Huet Stéphane
Jabaian Bassam
Jerbi Karim
Kahn Juliette
Kamiyama Takeki
Kern Sophie
King Hannah
Krzonowski Jennifer
Laaridh Imed

Labatut Vincent
Lalain Muriel
Lancien Mélanie
Larcher Anthony
Laurent Antoine
Lavoine Camille
Lazar Jan
Lefèvre Fabrice
Le Maguer Sébastien
Lemarchand Leslie
Liénard Jean-Sylvain
Luxardo Giancarlo
MacLeod A.N. Andrea
Magnen Cynthia
Mairano Paolo
Marczyk Anna
Martin Philippe
Mdhaffar Salima
Meignier Sylvain
Meunier Christine
Meynadier Yohann
Meziane Nacéra
Michaud Delfine
Michelas Amandine
Monnier Morgane
Nocaudie Olivier
Obin Nicolas
Paillereau Nikola
Pascal Pham
Pellegrino François
Pépiot Erwan
Perrier Pascal
Petitrenaud Simon
Philippart De Foy Marie
Piccaluga Myriam
Pillot-Loiseau Claire
Pinquier Julien
Pinto Serge
Popescu Anisia
Pouchoulin Gilles
Premat Timothée
Prince Typhanie
Pukli Monika
Rabant Stéphane
Rance Hélène
Raymond Michel
Rebourg Marie
Rezgui Dhouha
Ridouane Rachid
Riou Matthieu
Robert Danièle
Roebel Axel

Romeo Alice
Rosec Olivier
Rossato Solange
Rouvier Mickael
Roy Johanna-Pascale
Santiago Fabiàn
Savariaux Christophe
Schoentgen Jean
Schweitzer Claudia
Shao Bowei
Simonnet Edwin
Sock Rudolph
Solé Maria-Josep
Stam Gale
St-Gelais Xavier
Suire Alexandre
Tardieu Julien
Tellier Marion
Tomashenko Natalia
Tortel Anne
Tran Thi Thuy Hien
Trifu-Dejeu Loana
Turco Giuseppina
Vallée Nathalie
Vaxelaire Béatrice
Verhaegen Clémence
Vidailhet Marie
Vythelingum Kévin
Wang Ning
Woisard Virginie
Wottawa Jane
Yamaguchi Naomi
Yamlamai Nicha
Zaouali Hasna
Zhang Dan
Zhang Guoxian


## vendredi 8 juin - 9h30-13h :

### Atelier 1 - **PinPex, Patrimoine Instrumental de la Phonétique Expérimentale**

Le programme de l'atelier compte trois parties.
La première porte sur l'exposé par des représentants de cinq laboratoires francophones de leur collection instrumentale (et/ou documentaire) et des actions patrimoniales déjà menées ou en projet, relatives à l'inventaire, la restauration, l'exposition, la valorisation et les institutions partenaires.

La seconde propose un focus sur une pièce emblématique de chacune de ces collections ayant donné lieu à une action ou une valorisation particulière, par exemple une restauration, un documentaire, une exposition ou autre.

Enfin, après ce tour d'horizon, une discussion s'ouvrira sur les possibilités d'entreprendre des projets conjoints et fédératifs de sauvegarde et valorisation de ces collections et activités patrimoniales. Eparpillées, elles pourraient constituer un ensemble patrimonial de la communauté *Parole*, 'exposable' sur un espace muséal virtuel (web).

| | |
|---|---|
| 9h30-9h40 | Introduction de PinPex (Y. Meynadier & F. Hirsch) |
| 9h40-11h10 | **Collections et actions patrimoniales des laboratoires**<br>Aix-en- Provence (Y. Meynadier)<br>Montpellier (C. Dodane & F. Hirsch)<br>Belgique (B. Harmegnies)<br>Grenoble (C. Vilain)<br>Paris (D. Demolin)<br>Strasbourg (F. Hirsch, C. Fauth & B. Vaxelaire) |
| 11h10-11h30 | Pause-café |
| 11h30-12h20 | **Pièces emblématiques des laboratoires**<br>Le kymographe portable, Paris & Aix-en-Provence (A. Ghio, D. Demolin & Y. Meynadier)<br>Le segmentateur de Landercy, Mons (V. Delvaux)<br>L'analyseur de Koenig, Grenoble (C. Savariaux)<br>La collection ciné-radiographique de Straka, Strasbourg (B. Vaxelaire, F. Hirsch, P. Perrier & R. Sock)<br>Le photonasograph d'Ohala, Paris (A. Amelot) |
| 12h20-13h00 | **Discussion générale en table ronde**<br>Vers un projet fédératif de valorisation du patrimoine instrumental de phonétique expérimentale des laboratoires francophones de la Parole (Y. Meynadier & F. Hirsch) |

#### Organisateurs

**Yohann Meynadier**, Laboratoire Parole et Langage (LPL), CNRS – Aix-Marseille Université (AMU), Aix-en-Provence, yohann.meynadier@univ-amu.fr
**Fabrice Hirsch**, Laboratoire Praxiling, CNRS – Université Montpellier 3, Montpellier, fabrice.hirsch@univ-monpt3.fr

## Atelier 2 - **ExploSoc, Explorer les interactions sociales conversationnelles avec des agents artificiels**


Dans le domaine de la communication parlée, l'un des enjeux majeurs et interdisciplinaires aujourd'hui est de mieux comprendre les interactions sociales dans des situations écologiques, à l'échelle de la dyade ou du groupe. Les problématiques de recherche scientifiques sous-jacentes émanent de plusieurs disciplines : sciences de la parole et du langage, psychologie cognitive, informatique, neuroscience. Les agents artificiels communicants (en particulier les personnages virtuels ou les robots humanoïdes) peuvent être utilisés comme plateforme expérimentale pour mieux comprendre les mécanismes impliqués dans les interactions sociales qu'elles soient humain-humain ou humain-machine (Wykowska, Chaminade et Cheng, 2016).

En partant de cette constatation et dans le prolongement du workshop ISIA@ICMI 2017 (Chaminade, Nguyen, Ochs et Lefèvre 2017), le but de l'atelier est d'explorer comment les nouvelles technologies associées aux agents artificiels communicants peuvent être utilisées, ou le sont déjà, pour accroître notre connaissance des mécanismes des interactions sociales impliquant, par exemple les émotions, les signaux et fonctions sociales et la cognition. Des contextes d'interaction réalistes, en particulier utilisant le langage oral, nécessitent des outils et paradigmes expérimentaux nouveaux, combinant les sciences humaines et sociales avec les neurosciences et l'informatique, ainsi que de l'ingénierie. Cette vision globale et interdisciplinaire est nécessaire pour ouvrir la voie aux media sociaux du futur avec l'objectif d'améliorer la compétence sociale des agents artificiels, qu'il s'agisse de la compréhension mutuelle, la collaboration et le dialogue, avec des applications dans les domaines de la santé, de la formation ou des loisirs.

Wykowska, A., Chaminade, T., & Cheng, G. (2016). Embodied artificial agents for understanding human social cognition. Philos Trans R Soc Lond B Biol Sci, 371(1693). doi: 10.1098/rstb.2015.0375

Thierry Chaminade, Noël Ngyuen, Magalie Ochs, Fabrice Lefèvre: Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents, ISIAA@ICMI 2017, Glasgow, United Kingdom, November 13, 2017. ACM 2017, ISBN 978-1-4503-5558-2

Le programme de l'atelier compte quatre parties.

- **Session Présentation des enjeux** - Gérard Bailly du Gipsa-Lab, présentation des enjeux de l'approche [45']
- **Session Projets** - Présentation par les porteurs des avancées scientifiques réalisées dans des projets nationaux et internationaux en cours, dont H2020 CHIST-ERA JOKER, H2020 ANIMATAS, ANR IMPRESSIONS, ANR ACCORFORMED, PIA A*MiDex BLRI PSYSOCIAL [60']
- **Session Posters** - Pour les projets précédents et les participants, présentation des contributions scientifiques plus précises sous forme de poster [45']
- **Session Perspectives** - Discussion ouverte [15']

### Organisateurs

**Thierry Chaminade** (INT, ILCB, AMU), thierry.chaminade@univ-amu.fr
**Fabrice Lefèvre** (LIA, ILCB, U. Avignon), fabrice.lefevre@univ-avignon.fr
**Noël Nguyen** (LPL, ILCB, AMU), noel.nguyen@univ-amu.fr
**Magalie Ochs** (LIS, ILCB, AMU), magalie.ochs@lsis.org

## Atelier 3 - **Comment transposer les études en linguistique de l'oral en ressources pour l'enseignement du FLE**

Cet atelier se propose de traiter de la question de l'utilisation des corpus d'interactions authentiques comme ressources pertinentes pour l'enseignement de la compréhension de l'oral et de la gestion de l'interaction. Il sera présenté par des didacticiens et enseignants de FLE ainsi que des linguistes de l'oral.
Nous proposerons dans un premier temps des éléments méthodologiques permettant d'introduire les interactions naturelles en situation de classe. Puis, nous détaillerons des exemples d'exploitation des corpus oraux multimédia de FLEURON et CLAPI-FLE, et leurs retours d'expérience pour différents niveaux de langue et différents publics d'apprenants. Enfin, une session pratique permettra d'échanger autour des faits langagiers de l'oral que les participants souhaiteraient illustrer à partir de leurs propres besoins d'enseignement, vous pouvez nous les communiquer avant l'atelier : info-clapi@listes.ens-lyon.fr

9h30-10h **: Éléments méthodologiques permettant d'introduire les interactions naturelles en classe**

- Comment peut-on s'y prendre pour introduire puis travailler un extrait en compréhension de l'oral ou en interaction ?
- Quels paramètres prendre en compte à chaque niveau ? Quels enjeux ?

10h-10h30 : **Retour d'expérience sur l'introduction de certains extraits à des primo-apprenants**

- Est-il possible de sensibiliser des apprenants de FLE débutants à l'oral authentique ?
- Si oui quels sont les apports des corpus d'interactions authentiques par rapport à certains dialogues fabriqués proposés dans des manuels de FLE récents ? Présentation du dispositif et analyse de vidéos

10h30-11h00 : **Présentation de la ressource Fleuron et de son exploitation en classe**

- Le dispositif d'apprentissage en ligne Fleuron, les outils du site.
- Le corpus multimédia Fleuron : les données primaires et secondaires, les métadonnées, l'accès
- Exploitations didactiques : apprendre à interagir, faire de la (socio)linguistique de corpus.

11h00-11h15 : Pause

11h15-11h40 : **Présentation de la ressource CLAPI-FLE et de quelques exploitations**

- Les extraits : variation situationnelle, contextualisation, transcription modulable, exploitation
- Les collections : exemplier catégorisé et documenté
- Une palette d'utilisation pour différents niveaux d'apprenants

11h40-12h30 : **Session Pratique**

- Discussion et mise en pratique selon les demandes des participants : les salutations, la prise de rendez-vous, les questions, les atténuateurs, les reformulations, l'imparfait, les expressions, …

12h30-13h : **Retours et discussions**

### Organisateurs

**E. Oursel** (SFL, U. Paris 8), **C. Etienne** (ICAR, ENS Lyon), **E. Jouin-Chardon** (ICAR, ENS Lyon), **V. André** (ATILF, U. Lorraine), **C. David** (LPL, AMU)

## vendredi 8 juin - 14h-17h30 :

### Atelier 4 - **Regards croisés sur l'apraxie de la parole**

Au cours de ces dernières années, la communauté francophone a mis en place des projets qui s'attachent à l'étude de la parole apraxique, manifestant ainsi son intérêt pour ce trouble langagier encore mal appréhendé. L'apraxie de la parole est un trouble d'origine neurologique qui affecte le versant expressif du langage, notamment la production phonique des énoncés. Ce dysfonctionnement, souvent concomitant à l'aphasie de Broca, est une des séquelles les plus fréquentes d'un accident vasculaire cérébral chez l'adulte, mais il peut également constituer le premier signe des pathologies neurodégénératives, et en rester, pendant plusieurs années, la seule cause de handicap. Sur le plan cognitif, ce déficit est généralement attribué à l'interface entre l'encodage phonologique et l'exécution des gestes articulatoires, d'où son intérêt pour les sciences de la parole.
Nous partons du constat que l'étude de ce dysfonctionnement linguistique ne peut se faire sans faire correspondre les différentes dimensions de la parole apraxique :

i) sa dimension clinique, qui renvoie au diagnostic différentiel et à la prise en charge orthophonique
ii) sa dimension conceptuelle ou théorique : l'apraxie est une fenêtre sur l'organisation et le fonctionnement du langage chez le sujet sain et permet de tester la validité des concepts théoriques en linguistique (phonétique et phonologie), mais aussi en neuropsychologie
iii) sa dimension cérébrale, qui implique les bases anatomiques et fonctionnelles
iv) sa dimension expérimentale : l'étude de l'apraxie fait appel à des techniques d'investigation diverses et entraîne des problématiques méthodologiques.

Structurer ces multiples connaissances sur la parole apraxique constitue aujourd'hui un des enjeux majeurs dans la recherche sur la parole pathologique, notamment face au foisonnement d'études qui s'inscrivent dans ces différentes approches théoriques et méthodologiques. Pour répondre à cet enjeu, cet atelier a un double objectif. Plutôt que d'assembler des savoirs et des paradigmes difficilement comparables entre eux, nous proposerons de structurer la discussion à partir des problèmes fondamentaux qui s'attachent à l'étude de la parole apraxique. Par ailleurs, cet atelier visera à mobiliser les connaissances et les techniques d'investigation expérimentale des experts venant de disciplines voisines — pathologies du mouvement, linguistique dont la phonétique et la phonologie, sciences cognitives et neurosciences — autour de ces objectifs. La finalité de cette démarche est ainsi de tisser des liens entre ces disciplines, autour des problématiques communes se rapportant à la parole pathologique qui, à terme, permettront l'évolution d'une approche pluridisciplinaire vers une approche transdisciplinaire.
Le programme de l'atelier s'articule en fonction des objectifs évoqués ci-haut et prévoit 3 sessions majeures, à savoir :

**Data session** : exposé illustré par de multiples exemples de parole apraxique, par V. Sabadell et C. Verhaegen

**Table ronde** : structurée autour de problèmes fondamentaux présentés et contextualisés, avec la participation de Pascale Tremblay (CERVO), Rachide Ridouane (CNRS-LPP), Serge Pinto (CNRS-LPL), modération par C. Meunier.

**Perspectives** : débat ouvert, modéré par A. Marczyk et L. Baqué

#### Organisateurs

**A. Marczyk** (LPL, INS, ILCB, AMU), **V. Sabadell** (CHU Timone, ILCB, AMU), **C. Meunier** (LPL, ILCB, AMU), **A. Fasola** (INS-CNRS), **C. Verhaegen** (IRSTL, U. Mons), **L. Baqué** (FlexSem, U. Autonome Barcelone), **A. Rosas** (FlexSem , U. Autonome Barcelone), **T. Prince** (iBrain, INSERM, U. de Tours)

## Atelier 5 - **Réflexions pour le lancement d'un Groupe d'Instrumentation pour l'étude de la production et de la perception de la Parole**

Cet atelier vise à initier une réflexion au sein de la communauté Parole sur l'opportunité de créer un groupe d'instrumentation pour l'étude de la parole et du langage.

L'instrumentation dans le cadre de cet atelier s'entend au sens large ; il peut s'agir d'utilisation d'instrumentation existante, de conception et développement d'instruments scientifiques mais aussi de la réalisation de logiciels spécifiques qui peuvent être partagés au sein de la communauté. L'étude de la parole concerne les aspects dits « bas niveaux » mais aussi depuis plusieurs années les aspects dits « haut niveaux » avec le recours à l'électroencéphalographie (EEG et MEG) à l'IRM fonctionnelle (IRMf) et Fnir.

### Programme détaillé

| | |
|---|---|
| 14h-14h10 | Présentation de l'atelier |
| 14h10-14h15 | Tour d'horizon de la communauté |
| 14h15-14h30 | Présentation d'un retour d'expérience |
| 14h30-15h00 | Réflexions sur l'évolution de l'instrumentation (matérielle et logicielle) en parole à court et moyen termes |
| 15h00-15h30 | Formations : analyse de l'existant (pour qui, quel format, etc. ?) |
| 15h30-15h50 | Pause |
| 15h50-17h00 | Développement collaboratifs (matériel et logiciel) & Open science |
| 17h00-17h30 | Perspectives<br>· Organisation d'événements récurrents autour de l'instrumentation pour l'étude de la parole<br>· Répertoire des moyens (instruments, installations spécifiques) |

### Organisateurs

**Thibault Cattelain** (Gipsa Lab, U. Grenoble), thibault.cattelain@gipsa-lab.fr
**Thierry Legou** (LPL, ILCB, AMU), thierry.legou@lpl-aix.fr
**Fabrice Silva** (LMA, AMU), silva@lma.cnrs-mrs.fr

## Atelier 6 - **La parole dans l'espace de la classe : entre libération et didactisation**

Envisager la place que l'école accorde à la parole scolaire et à l'oralité, c'est s'interroger sur la manière dont la voix porte le langage en classe, que celui-ci soit outil de communication, moyen ou objet d'apprentissage.
Comment construire en situation didactique une scène verbale dans laquelle les voix de chacun, de l'enseignant comme des élèves, peuvent résonner conjointement ? Quelles tâches pour les élèves ?
Comment promouvoir l'expression individuelle en tenant compte de l'indispensable homogénéité d'un espace d'écoute privilégié collectif et consacré à l'acquisition des savoirs ? Quelle place donner à la voix physique, individuelle et singulière, à sa prosodie, sans qu'il y ait dissonance avec la voix des savoirs et du consensus didactique ?
Et peut-on tenir compte des modélisations existantes de la syntaxe de l'oral ?
Enfin, quels types de corpus constituer et à quelles fins ?

Ces problématiques seront abordées sous l'angle de l'enseignement du français langue de scolarisation et du français langue étrangère, et prendront appui sur des études menées dans des contextes d'enseignement français et suisses romands.

**14h-14h20 Introduction** : Quelles avenues d'enseignement de la parole scolaire : perspectives comparatistes suisses romandes et françaises ? *Véronique Bourhis et Roxane Gagnon*

**Volet A : Développement de la parole scolaire**

14h20-14h50 : D'une parole individuelle à une dynamique argumentative collective : un conseil d'élèves de 6e en REP. Quelle est la place des élèves ? *Sylvie Plane*

14h50-15h20 : L'oral scolaire, entre situations de communication pour développer l'oral, objets d'enseignement et oral pour apprendre. Une réflexion sur les ateliers philosophiques. *Joaquim Dolz*

**15h20 -15h30 Discussion à la suite du premier volet**

15h30 à 16h : Pause café

**Volet B : Quelle parole valoriser en classe de FL1/FLE ?**

16h-16h30 : Les commentaires métalangagiers évaluatifs des enseignants sur la parole de l'apprenant en FL1 et en FLE; multimodalité des pratiques d'évaluation. *Brahim Azaoui*

16h30-17h : Corpus de français oral annoté à des fins pédagogiques : les caractéristiques de l'oral en milieu universitaire dans l'enseignement/apprentissage du FLE. *Christian Surcouf et Alain Ausoni*

**17h-17h30 Synthèse et discussion à la suite de l'atelier ; échanges autour d'une éventuelle publication**

### Organisateurs

**Alain Ausoni**, *Université de Lausanne, alain.ausino@unil.ch*
**Brahim Azaoui**, *Université de Montpellier, brahim.azaoui@umontpellier.fr*
**Véronique Bourhis**, *Université de Cergy-Pontoise, veronique.bourhis@u-cergy.fr*
**Joaquim Dolz**, *Université de Genève, joaquim.dolz.mestre@unige.ch*
**Roxane Gagnon**, *HEP Vaud, roxane.gagnon@hepl.ch*
**Sylvie Plane**, *Sorbonne Université, sylvie.plane@wanadoo.fr*
**Christian Surcouf**, *Université de Lausanne, christian.surcouf@unil.ch*

Aix*Marseille université
Initiative d'excellence

cnrs

# JOURNEES D'ETUDES SUR LA PAROLE

# JEP 2018

LABORATOIRE PAROLE ET LANGAGE

32E EDITION AIX EN PROVENCE

## 4 JUIN

29 AVENUE ROBERT SCHUMAN

## 8 JUIN

AMPHITHEATRE 3

## ALLSH

JEP2018.SCIENCESCONF.ORG

UMR 7309 - CNRS - AMU PAROLE ET LANGAGE

AFCP

ISCA

EUROPEAN LANGUAGE RESOURCES ASSOCIATION ELRA

Ortolang — Open Resources and Tools for LANGuage

ILCB Institute of Language, Communication and the Brain

AIX MARSEILLE PROVENCE

FONDATION VASARELY