AFCP

ATALA

JEP-TALN-RECITAL 2016

Journées d'Études sur la Parole
Traitement Automatique des Langues Naturelles
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues

PARIS Inalco du 4 au 8 juillet 2016

Organisé par les laboratoires franciliens

https://jep-taln2016.limsi.fr

Conférenciers invités:

Christian Chiarcos (Goethe-Universität, Frankfurt.)
Mark Liberman (University of Pennsylvania, Philadelphia)

Coordinateurs comités d'organisation

Nicolas Audibert et Sophie Rosset (JEP)
Laurence Danlos & Thierry Hamon (TALN)
Damien Nouvel & Ilaine Wang (RECITAL)
Philippe Boula de Mareuil, Sarra El Ayari & Cyril Grouin (Ateliers)

Préface

Pour la cinquième fois, après Nancy en 2002, Fès en 2004, Avignon en 2008 et Grenoble en 2012, l'AFCP (Association Francophone pour la Communication Parlée) et l'ATALA (Association pour le Traitement Automatique des Langues) organisent conjointement leurs principales conférences pour réunir en un seul lieu les communautés du traitement des langues écrites, parlées et signées.

Plus précisément, la conférence JEP-TALN-RECITAL 2016 réunit cette année la 31e édition des Journées d'Étude sur la Parole (JEP 2016), la 23e édition de la conférence sur le Traitement Automatique des Langues Naturelles (TALN 2016) et la 18e édition des Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2016).

Cet événement est organisé par une quinzaine de laboratoires franciliens (ALPAGE, CLILLAC-ARP, ERTIM, IRCAM, LaTTiCe, LFF, LIPN, LIMSI, LPP, LSCP, LTCI, MoDyCo, PLIDAM, SFL, STIH) et se déroulera dans les locaux de l'INALCO (13e arrondissement de Paris).

JEP-TALN-RECITAL 2016 accueille deux conférenciers invités, Christian Chiarcos (Johann Wolfgang Goethe Universität Frankfurt a. M., Allemagne) et Mark Liberman (LDC & University of Pennsylvania, USA) ainsi que neuf ateliers :

— Celtic Language Technology Workshop (CLTW)
— Défi Fouille de Texte (DEFT)
— Enseignement des Langues et TAL (ELTAL)
— La Voix à la Barre
— Risque et TAL : détection, prévention et gestion
— Traitement automatique de la parole non standard
— Communautés en ligne : outils et applications en TAL (COLTAL)
— Traitement automatique des langues africaines (TALAf)
— Hackathon dans le domaine du TAL (HackaTAL)

Nous remercions tous les relecteurs et membres des différents comités de programme pour leur travail ainsi que nos sociétés savantes, l'AFCP et l'ATALA dont le CPERM (comité permanent) assure la continuité des éditions successives de TALN.

Pour les JEP, 99 articles ont été soumis, parmi lesquels 86 ont été sélectionnés, soit un taux de sélection de 87 %. 35 seront présentés en session orale et 51 lors de sessions posters.

Pour TALN, 41 articles longs ont été soumis, parmi lesquels 23 ont été sélectionnés, soit un taux de sélection de 56 %. 53 articles courts ont été soumis parmi lesquels 32 ont été sélectionnés pour un poster, soit un taux de sélection de 60 %.

Pour RECITAL, 11 articles ont été soumis, parmi lesquels 8 ont été sélectionnés, soit un taux de sélection de 73 %. Le faible nombre de soumissions reçues peut être, pour partie, lié au fait que de nombreux doctorants en première année de thèse soumettaient un article court co-signé par

leur(s) encadrant(s). Nous regrettons cette pratique qui prive les jeunes chercheurs des relectures pédagogiques, approfondies et signées. RECITAL est une conférence volontairement accessible, dans laquelle des travaux préliminaires peuvent être soumis et, pour certains, donner lieu à des présentations orales.

Nous rappelons que les conférences sont soucieuses du respect des règles de déontologie de la recherche et de la publication scientifique [1], qui s'imposent à tous. Les comités de programme sont donc particulièrement vigilants à cet égard.

Les actes de JEP-TALN-RECITAL-2016 sont en ligne sur le site `https://jep-taln2016.limsi.fr/actes`. Ils seront référencés par l'ACL (Association for Computational Linguistics) dans l'ACL Anthology (`http://www.aclweb.org/anthology/`).

Nous espérons que ces actes donneront à leurs lecteurs des idées fructueuses pour faire avancer la recherche et qu'ils serviront au rayonnement de la communauté francophone spécialiste du traitement des langues écrites, parlées et signées.

S. Rosset (LIMSI-CNRS)	N. Audibert (Univ. Paris 3, LPP-CNRS)	Présidents JEP
T. Hamon (Univ. Paris 13, LIMSI-CNRS)	L. Danlos (Univ. Paris-Diderot, ALPAGE-INRIA)	Présidents TALN
D. Nouvel (INALCO, ERTIM)	I. WANG (Université Paris 10, MoDyCo-CNRS)	Présidents RECITAL

1. Ces règles sont rappelées entre autres dans les chartes des thèses et par le COMETS, voir en particulier les pages 11 à 13 de ce document : `http://www.cnrs.fr/comets/IMG/pdf/guide_promouvoir_une_recherche_inte_gre_et_responsable_8septembre2014.pdf`.

Message des présidents de l'AFCP et de l'ATALA

C'est avec un plaisir toujours renouvelé que nous assistons à la tenue conjointe des Journées d'Études sur la Parole (JEP), de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) et des Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL). Ces conférences, issues de communautés scientifiques voisines, se sont retrouvées pour la première fois en 2002 à Nancy. Après le succès de la seconde organisation conjointe à Fès en 2004, l'ATALA et l'AFCP ont décidé d'une organisation commune tous les 4 ans. Cette cinquième édition conjointe — la 31ième édition des JEP, la 23ième de TALN et la 18ième de RECITAL — s'inscrit dans une série dont nous espérons la pérennité. L'organisation conjointe de ces événements se pare d'une aura particulière tant l'intersection entre communication parlée et traitement automatique des langues est grande. Le regard croisé sur l'écrit, l'oral et le signe, dans ses dimensions scientifiques, technologiques, applicatives et humaines, nous semble essentiel pour faire avancer l'état des connaissances sur ces différentes formes du langage, qui sont autant d'objets d'étude riches et complexes et pour renforcer la recherche francophone dans ces domaines.

Cet événement (car il faut avant tout voir cette réunion de JEP – TALN – RECITAL comme une entité à part entière et non comme trois manifestations dans un même lieu) est évidemment avant tout un événement scientifique pour échanger et faire le point sur l'état de l'art. Mais, de manière tout aussi importante, c'est également l'occasion de renforcer les liens sociaux entre chercheurs, entre communautés scientifiques, entre institutions académiques et industrielles et d'intégrer les étudiants et jeunes diplômés, qui souhaitent faire carrière dans nos domaines, au sein d'une dynamique collective pour leur faire profiter du capital de savoir et d'infrastructures accumulé au cours du temps. Préserver et développer des conférences francophones comme JEP, TALN et RECITAL est à notre avis vital pour la communauté, car ce sont d'abord des points d'entrée dans le monde de la publication scientifique où les jeunes chercheurs peuvent faire leur premiers pas, sans avoir à affronter la barrière de la langue. C'est aussi contribuer à la préservation de la diversité culturelle, tout en restant ouvert sur le monde. On peut même se prendre à rêver un jour de réunir dans un même lieu avec JEP, TALN et RECITAL les événements comparables qui existent dans les pays de langues romanes.

Construction d'une communauté scientifique, promotion d'une dynamique collective autour de la communication par le langage, et structuration des acteurs dans la sphère francophone nous apparaissent comme autant d'éléments importants pour résister à la dilution des thématiques scientifiques dans les applications et dans la course à l'impact sociétal dont nous sommes les témoins. Ce constat est particulièrement vrai dans nos disciplines où il est notamment exacerbé par la pluridisciplinarité inhérente à nos travaux. Nous savons tous à quel point mener des travaux mêlant plusieurs disciplines scientifiques est difficile, notamment en terme d'impact et de valorisation. De plus, la course actuelle au transfert industriel quasiment immédiat pour les disciplines scientifiques ayant des prolongements technologiques, et la quête de l'impact sociétal poussent à limiter la pluridisciplinarité à un simple assemblage de spécialités. Nous sommes persuadés que c'est par la discussion, l'échange et la fertilisation croisée entre les différentes disciplines en jeu dans l'étude de la communication parlée, écrite, ou signée, que nous pourrons faire émerger de nouveaux axes de recherche et permettre aux contributions fondamentales d'avoir la place qu'elles méritent dans le paysage de la recherche internationale. En particulier, le partage de ressources (données annotées mais aussi algorithmes,

protocoles d'évaluation, guide d'annotations, etc.) joue un rôle important dans la structuration de notre communauté et dans l'émergence de recherches pluridisciplinaires s'appuyant sur l'ensemble de nos champs d'investigation.

La tenue d'un événement joint est aussi l'occasion de s'interroger sur la manière dont nous abordons dans nos communautés respectives les changements, parfois révolutionnaires, qui surviennent dans notre pratique quotidienne de la recherche. Plusieurs évolutions sont venues récemment bouleverser nos disciplines en profondeur : les masses de données (big data) et son corollaire, l'informatique en nuage (cloud computing), ouvrent de nouvelles portes sur le traitement massif de données de parole et d'écrit ; l'apprentissage profond a investi le paysage scientifique, notamment avec l'avènement des modèles distributionnels ; et, moins récemment, des nouveaux médias et des médias sociaux ont émergés à l'échelle planétaire en un temps record, modifiant l'usage de la langue. Épiphénomènes ou modification profonde des fondements de nos domaines ? Le débat est ouvert et l'avenir le dira. En attendant, ces rencontres seront à n'en pas douter l'occasion de faire le point sur ces évolutions fortes, sur le plan scientifique bien sûr, mais aussi sur le plan de l'évolution générale de nos domaines. Le regard croisé de plusieurs disciplines, le partage d'expérience et de vision entre les différents domaines seront autant de richesses apportées à ce débat.

Il nous reste à clore ces quelques réflexions d'ouverture, en remerciant l'ensemble des personnes qui ont rendus possibles cet événement qui sera, à n'en pas douter, riche et passionnant. L'ATALA et l'AFCP tiennent tout d'abord à remercier les organisateurs des JEP, de TALN et de RECITAL pour l'énorme travail qu'ils ont réalisé afin d'assurer la qualité scientifique et la convivialité de l'événement. Nous constatons avec plaisir que tout a été mis en œuvre pour maximiser les échanges entre les différents acteurs du domaine, tous secteurs et tous domaines confondus. Nos remerciements vont également à l'ensemble des membres des comités de programme, sans qui une conférence ne serait qu'une longue litanie. Enfin, et surtout, les véritables garants de la qualité scientifique de la conférence sont les nombreuses personnes ayant participé à l'évaluation des soumissions, c'est-à-dire… la communauté scientifique ! Un grand merci aux relecteurs pour le temps qu'ils ont dédié à ce travail anonyme et pour la qualité du travail effectué qui se reflète dans le programme que vous aurez le plaisir de découvrir tout au long de la semaine.

Nous vous souhaitons une très bonne conférence, faite de découvertes scientifiques, d'échanges fructueux et de convivialité, en espérant que cette expérience vous donnera d'ores et déjà envie d'un prochain rendez-vous dans quatre ans.

Guillaume Gravier, président de l'Association Francophone de la Communication Parlée
Patrick Paroubek, président de l'Association pour le Traitement Automatique des LAngues

Comité d'organisation de JEP-TALN-RECITAL 2016 :

Martine Adda (CNRS, LPP)
Nicolas Audibert (Univ. Sorbonne Nouvelle, LPP)
Marion Blondel (CNRS, SFL)
Philippe Boula de Mareüil (CNRS, LIMSI)
Annelies Braffort (CNRS, LIMSI)
Ioana Chitoran (Univ. Paris Diderot, CLILLAC-ARP)
Chloé Clavel (Telecom-ParisTech, LTCI)
Alejandrina Cristia (CNRS, LSCP)
Laurence Danlos (Univ. Paris Diderot, ALPAGE)
Elisabeth Delais-Roussarie (CNRS, LLF)
Laurence Devillers (Univ. Paris-Sorbonne, LIMSI)
Sarra El Ayari (CNRS, SFL)
Michael Filhol (CNRS, LIMSI)
Karën Fort (Univ. Paris-Sorbonne, STIH)
Cécile Fougeron (Univ. Sorbonne Nouvelle, LPP)
Cyril Grouin (CNRS, LIMSI)
Thierry Hamon (Univ. Paris 13, LIMSI)
Agata Jackiewicz (Univ. Paris-Sorbonne, STIH)
Sylvain Kahane (Univ. Paris Ouest Nanterre, MoDyCo)
Jovan Kostov (INALCO, PLIDAM)
Thomas Lavergne (Univ. Paris-Sud, LIMSI)
Anaïs Lefeuvre (Univ. Paris-Sorbonne, STIH)
Anne-Laure Ligozat (ENSIIE, LIMSI)
Jean-Luc Minel (Univ. Paris Ouest Nanterre, MoDyCo)
Adeline Nazarenko (Univ. Paris 13, LIPN)
Aurélie Névéol (CNRS, LIMSI)
Damien Nouvel (INALCO, ERTIM)
Nicolas Obin (Univ. Pierre et Marie Curie, IRCAM)
Patrick Paroubek (CNRS, LIMSI)
Axel Roebel (Univ. Pierre et Marie Curie, IRCAM)
Sophie Rosset (CNRS, LIMSI)
Djamé Seddah (Univ. Paris-Sorbonne, ALPAGE)
Xavier Tannier (Univ. Paris-Sud, LIMSI)
Isabelle Tellier (Univ. Sorbonne Nouvelle, LaTTiCe)
Nadi Tomeh (Univ. Paris 13, LIPN)
Mathieu Valette (INALCO, ERTIM)
Cyril Verrecchia (CNRS, LIMSI)
Anne Vilnat (Univ. Paris-Sud, LIMSI)
Coralie Vincent (CNRS, SFL)
Ilaine Wang (Univ. Paris Ouest Nanterre, MoDyCo)
Pierre Zweigenbaum (CNRS, LIMSI)

Comité de programme de JEP 2016 :

Président :

Guillaume Gravier (CNRS, IRISA)

Membres :

Gilles Adda (CNRS, LIMSI)
Nicolas Audibert (Univ. Sorbonne Nouvelle, LPP)
Melissa Barkat-Defradas (Univ. Paul Valéry Montpellier 3, PRAXILING)
Jean-François Bonastre (Univ. d'Avignon et des Pays de Vaucluse, LIA)
Nathalie Camelin (Univ. du Maine, LIUM)
Alejandrina Cristia (CNRS, LSCP)
Elisabeth Delais-Roussarie (Univ. Paris Diderot, LLF)
Véronique Delvaux (FNRS/Univ. de Mons, Service de Métrologie et Sciences du Langage, Belgique)
Camille Fauth (Univ. de Strasbourg, LiLPa)
Benoit Favre (Aix-Marseille Univ., LIF)
Emmanuel Ferragne (Univ. Paris Diderot, CLILLAC-ARP)
Corinne Fredouille (Univ. d'Avignon et des Pays de Vaucluse, LIA)
Nathalie Henrich Bernardoni (CNRS, GIPSA-lab)
Frédéric Isel (Univ. Paris Ouest Nanterre, MoDyCo)
Juliette Kahn (LNE)
David Langlois (Univ. de Lorraine, LORIA)
Yohann Meynadier (Aix-Marseille Univ., LPL)
Claude Montacie (Univ. Paris-Sorbonne, STIH)
Nicolas Obin (Univ. Pierre et Marie Curie, IRCAM)
Thomas Pellegrini (Univ. Paul Sabatier Toulouse, IRIT)
Sophie Rosset (CNRS, LIMSI)
Sophie Wauquier (Univ. Paris 8, SFL)

Relecteurs additionnels :

Martine Adda-Decker (CNRS, LPP)
Moez Ajili (Université d'Avignon et des Pays de Vaucluse, LIA)
Vincent Arnaud (Université du Québec à Chicoutimi, Canada)
Grégoire Bachman (CNRS, LPP)
Pierre Badin (CNRS, GIPSA-lab)
Loïc Barrault (Université du Maine, LIUM)
Claude Barras (Université Paris-Sud, LIMSI)
Nathalie Bedoin (Université Lyon 2, DDL)
Frédéric Berthommier (Université Grenoble Alpes, GIPSA-lab)
Louis-Jean Boë (Université Grenoble Alpes, GIPSA-lab)
Anne Bonneau (CNRS, LORIA)
Stéphanie Borel (Université Sorbonne Nouvelle, LPP)
Fethi Bougares (Université du Maine, LIUM)
Philippe Boula De Mareüil (CNRS, LIMSI)
Damien Bouvier (CNRS, IRCAM)
Annelies Braffort (CNRS, LIMSI)
Mélanie Canault (Université Lyon 1, DDL)
Meunier Christine (CNRS, LPL)
Ioana Chitoran (Université Paris Diderot, CLILLAC-ARP)
Vincent Colotte (Université de Lorraine, LORIA)
Lise Crevier Buchman (CNRS, LPP)
Sébastien Delecraz (Aix-Marseille Université, LIF)
Paul Deléglise (Université du Maine, LIUM)
Didier Demolin (Université Sorbonne Nouvelle, LPP)
Giovanni Depau (Université Grenoble Alpes, GIPSA-lab)
Laurence Devillers (Université Paris-Sorbonne, LIMSI)
Mariapaola D'Império (Aix-Marseille Université, LPL)
Christelle Dodane (Université Paul Valéry Montpellier 3, PRAXILING)
Marion Dohen (Université Grenoble Alpes, GIPSA-lab)
David Doukhan (INA)
Richard Dufour (Université d'Avignon et des Pays de Vaucluse, LIA)
Camille Dutrey (LNE)
Yannick Estève (Université du Maine, LIUM)
Camille Fauth (Université de Strasbourg, LiLPa)
Isabelle Ferrané (Université Paul Sabatier Toulouse, IRIT)
Dominique Fohr (CNRS, LORIA)
Cécile Fougeron (Université Sorbonne Nouvelle, LPP)
Olivier Galibert (LNE)
Jiayin Gao (Université Sorbonne Nouvelle, LPP)
Silvia Gally (Université Grenoble Alpes, GIPSA-lab)
Maëva Garnier (CNRS, GIPSA-lab)
Cédric Gendrot (Université Sorbonne Nouvelle, LPP)
Laurianne Georgeton (Aix-Marseille Université, LPL)
Arseniy Gorin (CNRS, LIMSI)
James German (Aix-Marseille Université, LPL)
Alain Ghio (CNRS, LPL)
Anne Guyot Talbot (Université Paris Diderot, CLILLAC-ARP)
Bernard Harmegnies (Université de Mons, MSL, Belgique)
Sophie Herment (Aix-Marseille Université, LPL)
Thomas Hueber (CNRS, GIPSA-lab)
Céline Hidalgo (Aix-Marseille Université, LPL)
Fabrice Hirsch (Université Paul Valéry Montpellier 3, PRAXILING)
Kathy Huet (Université de Mons, MSL, Belgique)
Stéphane Huet (Université d'Avignon et des Pays de Vaucluse, LIA)
Irina Illina (Université de Lorraine, LORIA)
Bassam Jabaian (Université d'Avignon et des Pays de Vaucluse, LIA)
Philippe Joly (Université Paul Sabatier Toulouse, IRIT)
Denis Jouvet (INRIA, LORIA)
Takeki Kamiyama (Université Paris 8, DEPA, LPP)
Jennifer Krzonowski (CNRS, DDL)
Barbara Kuhnert (Université Sorbonne Nouvelle, LPP)
Anne Lacheret (Université Paris Ouest Nanterre, MoDyCo)
Muriel Lalain (CNRS, LPL)
Yves Laprie (CNRS, LORIA)
Anthony Larcher (I2R, A*STAR, Singapore)
Jérôme Lechien (Université de Mons, MSL, Belgique)
David Le Gac (Université de Rouen, Dysola)
Thierry Legou (CNRS, LPL)
Marco Liuni (IRCAM)
Marianne Louis (LPL)
Thibault Magallon (Aix-Marseille Université, LIF)
Sylvain Meignier (Université du Maine, LIUM)
Odile Mella (Université de Lorraine, LORIA)
Yohann Meynadier (Aix-Marseille Université, LPL)
Olivier Michalon (Aix-Marseille Université, LIF)
Alexis Michaud (CNRS, LACITO)
Julia Monnin (Université de la Nouvelle-Calédonie, CNEP)
Pascal Nocera (Université d'Avignon et des Pays de Vaucluse, LIA)
Francois Pellegrino (CNRS, DDL)
Pascal Perrier (Université Grenoble Alpes, GIPSA-lab)
Myriam Piccaluga (Université de Mons, MSL, Belgique)
Cristel Portes (Aix-Marseille Université, LPL)
Angélique Remacle (Université de Liège, ULV)
Rachid Ridouane (Université Sorbonne Nouvelle, LPP)
Albert Rilliard (CNRS, LIMSI)

Solange Rossato (Université Grenoble Alpes, LIG)
Amélie Rochet-Capellan (Université Grenoble Alpes, GIPSA-lab)
Marc Sato (CNRS, LPL)
Christophe Savariaux (Université Grenoble Alpes, GIPSA-lab)
Jean-Luc Schwartz (Université Grenoble Alpes, GIPSA-lab)
Christine Senac (Université Paul Sabatier Toulouse, IRIT)
Takaaki Shochi (Université Bordeaux Montaigne, CLLE ERSSàB)
Jérémie Tafforeau (Aix-Marseille Université, LIF)
Marie Tahon (IRISA)
Anne Tortel (Aix-Marseille Université, LPL)
Thi Thuy Hien Tran (Université Stendhal, GIPSA-lab)
Jérémy Trione (Aix-Marseille Université, LIF)
Gabor Turcsan (Aix-Marseille Université, LPL)
Nathalie Vallee (Université Grenoble Alpes, GIPSA-lab)
Ioana Vasilescu (CNRS, LIMSI)
Christophe Veaux (IRCAM)
Coriandre Vilain (Université Grenoble Alpes, GIPSA-lab)
Emmanuel Vincent (INRIA, LORIA)
Hiyon Yoo (Université Paris Diderot, LLF)

Comité de programme de TALN 2016 :

Présidents :

Laurence Danlos (Univ. Paris-Diderot, ALPAGE)

Thierry Hamon (Univ. Paris 13, LIMSI)

Membres du comité de programme :

Laurent Besacier (Univ. de Grenoble, LIG)
Vincent Claveau (CNRS, IRISA)
Olivier Ferret (CEA LIST)
Laurence Danlos (Univ. Paris Diderot, ALPAGE)
Gaël Dias (Univ. de Caen Basse-Normandie, GREYC)
Thierry Hamon (Univ. Paris 13, LIMSI)
Nabil Hathout (CNRS, CLLE)
Philippe Langlais (Univ. de Montréal, RALI, Canada)
Laurence Meurant (Univ. de Namur, LSFB, Belgique)
Emmanuel Morin (Univ. de Nantes, LINA)
Adeline Nazarenko (Univ. Paris 13, LIPN)
Pascale Sébillot (INSA de Rennes, IRISA)

Membres du comité de relecture :

Stergos Afantenos (Univ. Paul Sabatier, IRIT)
Salah Aït-Mokhtar (Xerox Research Centre Europe)
Maxime Amblard (Univ. de Lorraine, LORIA)
Jean-Yves Antoine (Univ. François Rabelais Tours/Blois, LI)
Delphine Battistelli (Univ. Paris Ouest Nanterre, MoDyCo, CNRS)
Frédéric Béchet (Aix-Marseille Univ., LIF)
Delphine Bernhard (Univ. de Strasbourg, LiLPa)
Romaric Besançon (CEA, LIST)
Brigitte Bigi (CNRS, LPL)
Philippe Blache (CNRS, LPL)
Hervé Blanchon (Univ. Pierre-Mendès-France, LIG)
Marion Blondel (Univ. Paris 8, SFL)
Florian Boudin (Univ. de Nantes, LINA)
Annelies Braffort (CNRS, LIMSI)
Nathalie Camelin (Univ. du Maine, LIUM)
Thierry Charnois (Univ. Paris 13, LIPN)
Guillaume Cleuziou (Univ. d'Orléans, LIFO)
Benoit Crabbé (Univ. Paris Diderot, ALPAGE)
Beatrice Daille (Univ. de Nantes, LINA)
Marco Dinarelli (CNRS, LaTTiCe)
Iris Eshkol (Univ. d'Orléans, LLL)
Yannick Estève (Univ. du Maine, LIUM)
Cécile Fabre (Univ. Toulouse - Jean Jaurès, CLLE)
Benoit Favre (Aix-Marseille Univ., LIF)
Karën Fort (Univ. Paris-Sorbonne, STIH)
Thomas François (Univ. Catholique de Louvain, CENTAL, Belgique)
Nathalie Friburger (Univ. François Rabelais, LI)
Éric Gaussier (Univ. Joseph Fourier, LIG)
Natalia Grabar (CNRS, STL)
Lamia Hadrich Belguith (Univ. de Sfax, MIRACL, Tunisie)

Nicolas Hernandez (Univ. de Nantes, LINA)
Stéphane Huet (Univ. d'Avignon et des Pays de Vaucluse, LIA)
Sylvain Kahane (Univ. Paris Ouest Nanterre, MoDyCo)
Olivier Kraif (Univ. Stendhal Grenoble 3, LIDILEM)
Mathieu Lafourcade (Univ. de Montpellier, LIRMM)
Guy Lapalme (Univ. de Montréal, Canada)
Joseph Le Roux (Univ. Paris 13, LIPN)
Jean-Marc Lecarpentier (Univ. de Caen Basse-Normandie, GREYC)
Anne-Laure Ligozat (ENSIIE, LIMSI)
Denis Maurel (Univ. Francois Rabelais Tours, LI)
Richard Moot (CNRS, LaBRI)
Véronique Moriceau (Univ. Paris-Sud, LIMSI)
Philippe Muller (Univ. Paul Sabatier Toulouse, IRIT)
Jian-Yun Nie (Univ. de Montréal, Canada)
Aurélie Névéol (CNRS, LIMSI)
Yannick Parmentier (Universite d'Orléans, LIFO)
Thierry Poibeau (CNRS, LaTTiCe)
Andrei Popescu-Belis (IDIAP Research Institute, Suisse)
Jean-Philippe Prost (Univ. de Montpellier, LIRMM)
Solen Quiniou (Univ. de Nantes, LINA)
Christian Raymond (INSA de Rennes, IRISA)
Christian Retoré (Univ. de Montpellier, LIRMM)
Mathieu Roche (Cirad, TETIS)
Didier Schwab (Univ. Pierre-Mendès-France, LIG)
Djamé Seddah (Univ. Paris-Sorbonne, ALPAGE)
Kamel Smaïli (Univ. de Lorraine, LORIA)
Xavier Tannier (Univ. Paris-Sud, LIMSI)
Isabelle Tellier (Univ. Sorbonne Nouvelle, LaTTiCe)
Juan-Manuel Torres-Moreno (Univ. d'Avignon et des Pays de Vaucluse, LIA)
Christel Vrain (Univ. d'Orléans, LIFO)
Eric Wehrli (Univ. de Genève, Suisse)
Guillaume Wisniewski (Univ. Paris-Sud, LIMSI)
François Yvon (Univ. Paris-Sud, LIMSI)
Pierre Zweigenbaum (CNRS, LIMSI)

Relecteurs additionnels :

Chafik Aloulou (Univ. de Sfax, MIRACL, Tunisie)
Ibrahim Bounhas (Univ. de la Manouba, LISI, Tunisie)
Eric De La Clergerie (INRIA, ALPAGE)
Maher Jaoua (Univ. de Sfax, MIRACL, Tunisie)
Hugo Mougard (Univ. de Nantes, LINA)
Yuliya Korenchuk (Univ. de Strasbourg, LiLPa)
Ophélie Lacroix (CNRS, LIMSI)
Souha Mezghani (Univ. de Sfax, MIRACL, Tunisie)
Sarah Zenasni (Univ. de Montpellier, LIRMM)

Comité de programme de RECITAL 2016 :

Présidents :

Damien Nouvel (Inalco, ERTIM)
Ilaine Wang (Univ. Paris 10, MoDyCo)

Membres :

Maxime Amblard (Univ. de Lorraine, LORIA)
Jean-Yves Antoine (Univ. François Rabelais Tours/Blois, LI)
Patrice Bellot (Polytech Marseille - Aix-Marseille Univ., LSIS)
Houda Bouamor (Carnegie Mellon University, Qatar)
Florian Boudin (Univ. de Nantes, LINA)
Annelies Braffort (CNRS, LIMSI)
Sandra Bringay (Univ. Paul Valéry Montpellier 3, LIRMM)
Marie Candito (Univ. Paris Diderot, ALPAGE)
Gaël de Chalendar (CEA, LIST)
Marcel Cori (Univ. Paris Ouest Nanterre, MoDyCo)
Benjamin Duthil (Univ. de La Rochelle, L3I)
Maud Ehrmann (EPFL, Suisse)
Benoit Favre (Aix-Marseille Univ., LIF)
Karën Fort (Univ. Paris-Sorbonne, STIH)
Thomas François (Univ. Catholique de Louvain, CENTAL, Belgique)
Bruno Gaume (Univ. Toulouse - Jean Jaurès, CLLE)
Mathieu Lafourcade (Univ. de Montpellier, LIRMM)
Philippe Langlais (Univ. de Montréal, RALI, Canada)
Charlotte Lecluze (Univ. de Caen Basse-Normandie, GREYC)
Benjamin Lecouteux (Univ. Pierre-Mendès-France, LIG)
Anaïs Lefeuvre (Univ. Paris-Sorbonne, STIH)
Gaël Lejeune (Univ. de Caen Basse-Normandie, GREYC)

Coordinateurs ateliers :

Soutiens

Partenaires institutionnels

Entreprises

JEP-TALN-RECITAL 2016 est soutenu par Google

Table des matières

Le VOT des éjectives : le cas du maya yucatèque

Emre Bayraktar [1] Rachid Ridouane [1]
(1) Laboratoire de Phonétique et Phonologie, (CNRS / Sorbonne Nouvelle), 19 Rue des Bernardins, 75005 Paris, France
emre.bayraktar@univ-paris3.fr, rachid.ridouane@univ-paris3.fr

RÉSUMÉ

Cet article présente une étude acoustique des occlusives éjectives du maya yucatèque. S'intéressant spécifiquement au voice onset time (VOT), l'étude examine d'une part si le VOT est un corrélat acoustique fiable de l'éjectivité dans cette langue et d'autre part si le VOT varie selon le lieu d'articulation et la hauteur vocalique. Les résultats, obtenus à partir des productions de deux locuteurs natifs, montrent que les éjectives ont un VOT plus long comparées à leurs contreparties pulmonaires. Parmi les éjectives, le VOT varie en fonction du lieu d'articulation, les vélaires présentant le VOT le plus long. De même une tendance pour un VOT plus court devant les voyelles hautes a été observée. Ces résultats soulèvent un ensemble de questions concernant les mécanismes qui sous-tendent les variations du VOT, notamment en lien avec les contraintes aérodynamiques en jeu lors de la production des occlusives éjectives.

ABSTRACT

The VOT of ejective stops in Maya Yucatec
This article presents an acoustic study of ejective stops in Maya Yucatec. Focusing on voice onset time (VOT), it examines, on the one hand, whether VOT is a reliable acoustic correlate to ejectivity and, on the other hand, whether VOT varies depending on place of articulation and vowel height. Results, obtained from the productions of two native speakers, show that ejectives have longer VOTs than their pulmonic counterparts. Within ejectives, VOT varies depending on place of articulation of stops, VOT for velars being the longest. VOT of ejectives also tends to be shorter before high vowels. These results raise a set of questions concerning the mechanisms underlying VOT variations. In particular, we consider the possible aerodynamic constraints at play during the production of ejective stops.

MOTS-CLÉS : éjectives, maya yucatèque, VOT, lieu d'articulation, hauteur vocalique.
KEYWORDS: ejectives, Maya Yucatec, VOT, place of articulation, vowel height.

1 Introduction

Le maya yucatèque fait partie de la famille maya. Cette famille comprend plus d'une vingtaine de langues, parlées notamment au Guatemala, au Belize et au Mexique (Colazo-Simon, 2007). Le

maya yucatèque est parlé essentiellement au Yucatán (un État situé au sud-est du Mexique, sur la péninsule du Yucatán), ainsi qu'au Belize. Cette langue est parlée par 700 000 locuteurs environ (Lewis, 2009). Son inventaire phonémique comporte 20 consonnes, dont cinq éjectives. Trois sont des occlusives et deux sont des affriquées (voir table 1).

		labiale	alvéolaire	post-alvéolaire	palatale	vélaire	glottale
occlusive	pulmonaire	p	t			k	ʔ
	éjective	p'	t'			k'	
	implosive	ɓ					
affriquée	pulmonaire		ts	tʃ			
	éjective		ts'	tʃ'			
fricative			s	ʃ			h
nasale		m	n				
approximante latérale			l				
approximante		w			j		

TABLE 1 : Inventaire des consonnes du maya yucatèque (Bricker et al., 1998)

Les éjectives sont des phonèmes qui s'opposent à leurs contreparties pulmonaires, comme en atteste l'existence de nombreuses paires minimales (e.g. /*kan*/ 'quatre' vs. /*k'an*/ 'jaune' ; /*kool*/ 'milpa' vs. /*k'ool*/ 'frapper' ; /*tíit*/ 'remuer' vs. /*t'íit*/ 'répandre' ; /*tuup*/ 'boucle d'oreille' vs. /*t'uup*/ 'petit doigt' ; /*paak*/ 'plier' vs. /*p'aak*/ 'tomate' ; /*péek*/ 'bouger' vs. /*p'eek*/ 'haïr').

1.1 Production des éjectives : bref aperçu

La production d'une occlusive éjective se caractérise par un mécanisme de pression d'air non pulmonaire. La glotte est fermée pour isoler la cavité orale de la cavité sous-glottique. Le larynx s'élève et fait augmenter la pression supraglottique. C'est cette pression que le locuteur utilise pour produire la perturbation du flux d'air au moment du relâchement (Catford, 1939). La cavité orale subit une contraction maximale afin d'y comprimer l'air présent. Cette compression provoque un fort burst lors du relâchement, et le voisement se fait tardivement car les plis vocaux restent fermement serrés un certain moment après le relâchement de l'occlusion buccale et au-dessus de leur position habituelle, rendant ainsi l'initiation de leurs vibrations plus tardive (Ladefoged & Maddieson, 1996). Ce retard de voisement se caractérise généralement par une période de silence entre le relâchement de l'occlusive et l'onset de la voyelle qui suit. Cette caractéristique a été observée pour le tigrinya (Kingston, 1985) ou encore l'ingush (Warner, 1996). Nous avons observé cette période de silence (ou quasi-silence) en maya yucatèque (voir figures 1 et 2) chez les deux locuteurs de notre étude (de manière systématique pour le locuteur 1 et majoritaire pour le locuteur 2).

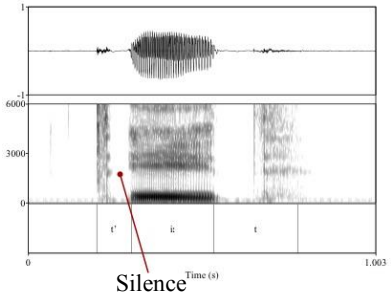


FIGURE 1 : Ejective coronale [t'] dans le mot [t'iit] 'répandre', loc. 1

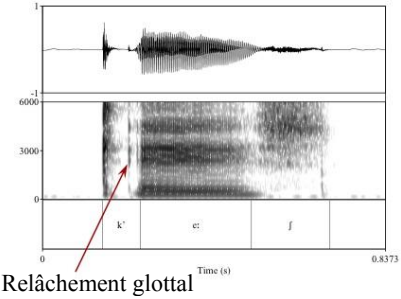


FIGURE 2 : Ejective vélaire [k'] dans le mot [k'eex] 'réponse', loc. 2

Sur certaines productions, les conséquences acoustiques d'un coup de glotte sont visibles vers la fin de la période de silence (voir figure 2). Le voice onset time (VOT) des éjectives du maya yucatèque est donc composé de deux périodes, une période de bruit de relâchement, suivie d'une période de (quasi-) silence. Une conséquence possible au retard de voisement observé pour les éjectives serait un allongement de la durée du VOT.

1.2 VOT des éjectives et ses variations

Le VOT, qui est généralement compris comme l'intervalle temporel entre le relâchement oral et l'onset du voisement de la voyelle qui suit, est un corrélat acoustique majeur permettant de caractériser les occlusives selon leurs traits laryngaux (voir notamment Lisker & Abramson (1964) et son travail pionnier sur les occlusives voisées et non voisées de 11 langues). La durée du VOT permet en effet de catégoriser différentes séries d'occlusives, comme [+ voisé], [-voisé] ou [+ aspiré] (Cho & Ladefoged, 1999 ; Wysocki, 2004). Il ressort de la revue de la littérature que les éjectives ont généralement un VOT plus long que leurs contreparties pulmonaires (Lindau, 1984). C'est le cas notamment en tlingit, où les occlusives pulmonaires ont un VOT de 24,6 ms en moyenne, tandis que les occlusives éjectives ont un VOT de 102,7 ms (Maddieson et al., 2001). Pour autant cette tendance pour un VOT plus long pour les éjectives n'a pas été observée pour toutes les langues. Par exemple, dans l'ingush, le VOT des éjectives est seulement de 26,2 ms en moyenne, alors qu'il est de 45 ms pour les pulmonaires (Warner, 1996).

Il est très largement admis que la durée du VOT varie selon le lieu d'articulation des occlusives pulmonaires (Cho & Ladefoged, 1999). Pour les éjectives, les données dans la littérature ne montrent pas d'effet aussi systématique. Cho & Ladefoged (1999) ont souligné ce pattern irrégulier, qu'ils associent aux différences entre les langues dans le degré de l'élévation du larynx ou dans le timing entre le relâchement oral et le relâchement glottal. Ils soutiennent ainsi l'absence de mécanisme physiologique permettant de rendre compte de l'effet de lieu sur le VOT dans la réalisation des éjectives. Il a été noté, cependant, que plusieurs langues non apparentées font état de différences significatives de VOT des éjectives en fonction du lieu d'articulation. C'est le cas notamment pour le yapese et le nez perce (Maddieson, 2001).

Une autre source de variation du VOT des occlusives pulmonaires concerne le contexte vocalique. Les résultats les plus communément rapportés dans la littérature mettent en lumière une

différence de VOT de l'occlusive en fonction de la hauteur de la voyelle qui suit : les voyelles hautes provoquent un VOT plus long que les voyelles basses (Weismer, 1979 ; Higgins et al., 1998). A notre connaissance, il n'a jamais été relevé dans la littérature d'effet de la hauteur de la voyelle sur le VOT des éjectives.

1.3 Problématique

A la lumière de cette revue de la littérature, plusieurs questions se posent. Une première question est de savoir si le VOT permet de distinguer les occlusives éjectives du maya yucatèque de leurs contreparties pulmonaires. Malgré quelques contre-exemples dans certaines langues, nous nous attendons à observer pour le maya yucatèque un VOT plus long pour les éjectives que pour les occlusives pulmonaires. Ce VOT plus long pour les éjectives serait une conséquence acoustique de l'articulation qui nécessite une période supplémentaire pour abaisser le larynx à sa position optimale pour le voisement. Une deuxième question concerne l'effet de lieu d'articulation sur la durée du VOT. Allons-nous observer le même pattern pour les occlusives éjectives ? Autrement dit, est-ce que les éjectives vélaires présentent des VOT plus longs que les éjectives labiales et coronales en maya yucatèque ? Pour rappel, Cho & Ladefoged (1999) ayant effectué une étude similaire sur 6 langues comportant des éjectives, n'avaient trouvé aucune différence significative du VOT des éjectives selon le lieu d'articulation. Nous nous attendons à observer les mêmes résultats pour le VOT des éjectives du maya yucatèque. La troisième question concerne l'effet des voyelles post-occlusives sur la durée du VOT, en fonction de la hauteur de la langue. Nous savons que la durée du VOT des occlusives pulmonaires varie selon ce paramètre. Qu'en est-il des éjectives ? Notre hypothèse est que le même pattern sera observé, c'est-à-dire un VOT plus long des éjectives devant la voyelle haute [i], et un VOT plus court devant la voyelle basse [a]. Pour les occlusives pulmonaires, cette variation du VOT est liée à la position de la langue lors de la production de la voyelle suivante. Au moment du relâchement de l'occlusive, si la langue est en position haute, l'air mettra plus de temps à être évacué du conduit vocal que si la langue est en position basse (Chang et al., 1999). Il serait logique que ces mêmes contraintes s'opèrent lors de la production d'une éjective.

2 Méthode

Les données acoustiques ont été enregistrées auprès de deux locuteurs natifs du maya yucatèque. Un locuteur masculin (loc. 1), de 45 ans, originaire de Mérida, au Mexique et une locutrice (loc. 2) âgée de 32 ans et originaire de Tlalpan, au Mexique. Les deux locuteurs sont bilingues (parlant aussi l'espagnol).

Notre étude porte sur les occlusives éjectives (/p'/, /t'/ et /k'/) et leurs contreparties pulmonaires (/p/, /t/ et /k/), en position initiale de mot, suivies des /a/, /e/, /i/, /o/, /u/, représentant l'inventaire vocalique du maya yucatèque. Contraints par la distribution des consonnes et voyelles de la langue, nous n'avons pu recueillir de mots avec la voyelle /u/ après les vélaires /k/ et /k'/, la voyelle /o/ après la coronale /t/, ainsi que la voyelle /e/ après la vélaire /k/. Le corpus analysé est constitué de 26 mots (20 mots monosyllabiques, 6 mots bisyllabiques), chaque mot étant répété 12 fois par chacun des deux locuteurs (6 répétitions à l'isolé, et 6 répétitions en phrase cadre). Lors de l'acquisition des données, des omissions involontaires et autres incidents d'enregistrement ont

modifié le nombre initial des occurrences pour le locuteur 1. Les occurrences de ce locuteur représentent ainsi un total de 188 items, constitués de 81 occlusives pulmonaires (36 [p], 25 [t] et 20 [k]) et 107 occlusives éjectives (44 [p'], 38 [t'] et 25 [k']). Les occurrences en contexte [ta], [ki] et [k'a] de ce locuteur ont été exclues pour l'étude sur la hauteur vocalique, par manque d'effectif suffisant. Le corpus du locuteur 2 représente 312 items. Les données analysées pour ce locuteur sont constituées de 144 occlusives pulmonaires (60 [p], 48 [t], 36 [k], et 168 occlusives éjectives (60 [p'], 60 [t'] et 48 [k']).

Pour le locuteur 1, l'enregistrement s'est effectué dans la chambre sourde du Laboratoire de Phonétique et Phonologie (CNRS/Sorbonne Nouvelle) à Paris, à l'aide d'un micro-casque hypercardioïde (AKG C520), une carte son Edirol UA25 et le logiciel Sound Studio, v. 3.6.0.0. L'enregistrement du locuteur 2 a aussi été effectué dans une chambre sourde mais à l'Université de Bielefeld en Allemagne, à l'aide d'un microphone à capsule large cardioïde (Neumann TLM 103), une carte son Pro Tools, v. 8.1 qui possède son propre logiciel d'enregistrement.

Les données acoustiques on été segmentées, annotées et analysées à l'aide de Praat (Boersma & Weenink 2012). Comme nous l'avons signalé plus haut, le VOT est compris ici comme l'intervalle temporel entre l'onset du relâchement oral et l'onset du voisement de la voyelle qui suit.

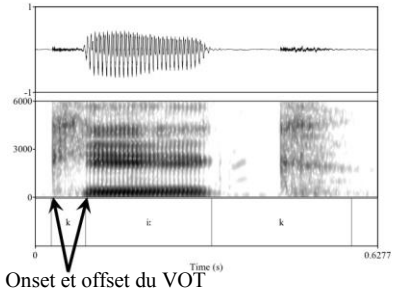


FIGURE 3 : VOT de la pulmonaire [k]

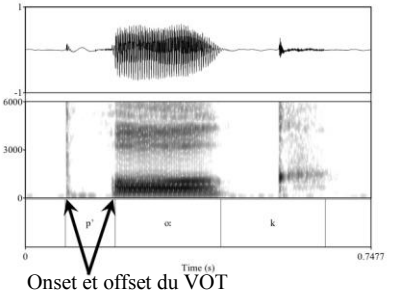


FIGURE 4 : VOT de l'éjective [p']

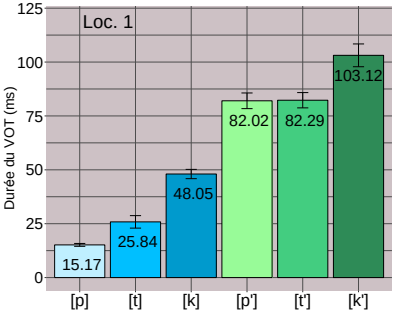
L'onset du relâchement correspond à une apparition abrupte d'énergie à certaines hauteurs fréquentielles correspondant aux formants, ou plus précisément au F-pattern de l'occlusive, et l'onset du voisement correspond à la première des stries verticales régulières, indiquant la pulsation glottale, illustrant la fin du VOT (voir figure 3). Pour les éjectives, la mesure du VOT est la même, nous délimitons la fin du VOT aux premières vibrations des plis vocaux (voir figure 4). Cette méthode s'aligne, entre autres, sur celles de Maddieson (2001), Wright et al. (2002). Notre étude ne prétend pas fournir de résultats statistiques robustes car nous avons effectué nos tests sur une petite population. Il s'agit donc de rester prudent quant à la représentativité des résultats sur cette langue. Puisque nous possédons des effectifs qui ne sont pas systématiquement identiques en nombres, et qui ne sont pas élevés de surcroît (environ quelques dizaines lors de chaque étude comparative), la prudence nous pousse à choisir des tests non-paramétriques pour observer la significativité des différents effets étudiés. Quand il s'agit de comparer deux groupes d'occlusives (par exemple pulmonaires vs. éjectives), nous utiliserons le test U Mann-Whitney. Quand il s'agit de comparer trois groupes d'occlusives (par exemple /p'/ vs. /t'/ vs. /k'/), nous utiliserons le test Kruskal-Wallis.

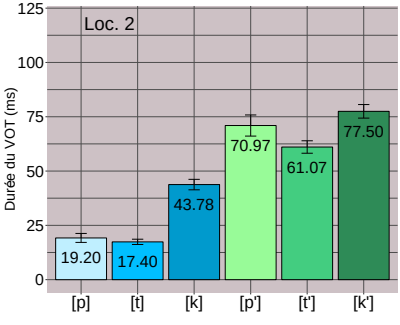
Pour chaque test, la valeur de « p » mentionnée représente le taux de significativité. Comme nous l'avons indiqué plus haut, notre étude est limitée aux occlusives en position initiale de mot, car cette position est celle généralement utilisée dans la littérature (Lisker & Abramson, 1964 ; Cho & Ladefoged, 1999). Cela nous permettra ainsi d'établir plus facilement des comparaisons avec les résultats d'autres auteurs.

3 Résultats

3.1 VOT des éjectives et effet de lieu d'articulation

Les résultats montrent que les occlusives éjectives ont un VOT significativement plus long comparées à leurs contreparties pulmonaires ($p < 0{,}0001$), et ce pour chaque locuteur et chaque lieu d'articulation (voir figures 5 et 6 ; les barres représentent les intervalles de confiance).





FIGURES 5 ET 6 : VOT des différentes occlusives pour chaque locuteur

Pour les occlusives pulmonaires, le test de Kruskal-Wallis montre que les deux locuteurs ont des VOT significativement différents selon le lieu d'articulation. Les vélaires ayant un VOT plus long que les labiales et les coronales (locuteur 1 : $\chi 2\ (2) = 54.728$, $p < 0.0001$, locuteur 2 : $\chi 2\ (2) = 59{,}442$, $p < 0.0001$). Dans le cas des éjectives, les deux locuteurs ont aussi un VOT de l'éjective vélaire significativement plus long que les occlusives alvéolaires et bilabiales (locuteur 1 : $\chi 2\ (2) = 11.226$, $p < 0.01$, locuteur 2 : $\chi 2\ (2) = 12.427$, $p < 0.01$). Ce résultat indique ainsi que, à l'instar des occlusives pulmonaires, le VOT des éjectives varie selon le lieu d'articulation des consonnes. Pour autant, aucune différence significative n'a été observée entre les labiales et les coronales chez les deux locuteurs.

3.2 VOT des éjectives selon la hauteur de la voyelle suivante

A notre connaissance, il n'a jamais été relevé dans la littérature d'effet de la hauteur de la voyelle sur le VOT des éjectives. Nous avons constaté des résultats similaires dans nos données, à l'exception d'un cas pour le locuteur 1, dont le VOT de [p'] est significativement plus court devant [i], comparé à [a] ($p = 0{,}0134$). A la vue des différentes moyennes de VOT selon la hauteur vocalique (voir figure 7), une tendance se dégage : les voyelles basses semblent induire un VOT plus long que les voyelles hautes. Ces résultats sont intrigants car ils sont à l'opposé des résultats

observés pour les occlusives non-éjectives. A noter néanmoins que pour les deux locuteurs, les résultats pour [k'] sont similaires à ceux rapportés pour les occlusives non-éjectives (i.e. un VOT plus long devant [i]).

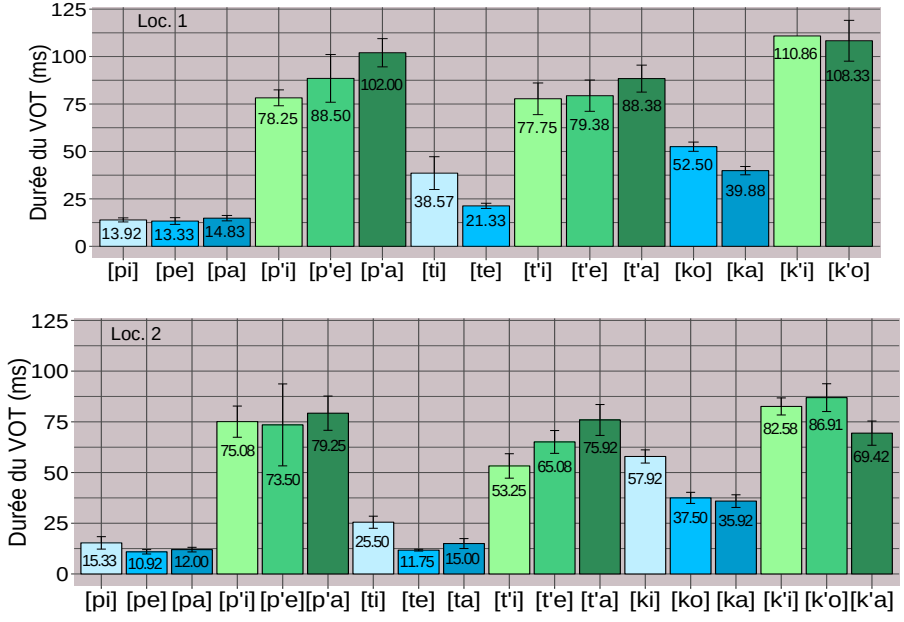


FIGURE 7 : VOT des différentes occlusives selon la hauteur vocalique pour chaque locuteur

4 Discussion

Dans notre étude, nous avons montré qu'en maya yucatèque le VOT des éjectives est significativement plus long que le VOT des pulmonaires. Les trois lieux d'articulation observés montrent chacun un VOT significativement plus long que sa contrepartie pulmonaire. Ce résultat rejoint la plupart des résultats sur les éjectives dans d'autres langues ; notamment, comme pour le maya yucatèque, lorsque les éjectives sont comparées aux pulmonaires non aspirées. Nous pouvons donc avancer que le VOT est un corrélat important permettant de distinguer acoustiquement les éjectives du maya yucatèque de leurs contreparties pulmonaires. A l'évidence cette différence n'est pas uniquement de nature temporelle. Comme nous l'avons signalé plus haut, une caractéristique importante des éjectives est la phase de (quasi)-silence contenu dans cet intervalle temporel, correspondant à la phase d'occlusion glottale après le relâchement oral. Phase durant laquelle le larynx s'abaisse et les plis vocaux commencent à se relâcher. Cette étude a aussi montré que les éjectives vélaires avaient un VOT plus long que les éjectives coronales et labiales. Ce résultat montre un pattern similaire à ce qui a été observé pour les occlusives pulmonaires, suggérant la possible existence d'une explication unifiée à ce phénomène. Pour autant, si les raisons expliquant ce phénomène sont assez bien connues pour les pulmonaires, elles le sont moins pour les éjectives. En effet, pour les pulmonaires, ces variations peuvent être dues à un ensemble de facteurs, de nature aérodynamique ou articulatoire (voir Cho & Ladefoged, 1999 pour une revue détaillée). Des

facteurs tels que la taille de la cavité derrière le point de constriction, la vélocité des articulateurs, ou encore l'étendue du contact articulatoire.

Maddieson (2001) propose une piste pour expliquer l'effet de lieu sur la durée du VOT des éjectives. Il reprend l'observation selon laquelle la phase d'occlusion est inversement liée à la période de relâchement, et avance l'hypothèse qu'un VOT plus long suivant une phase d'occlusion courte pourrait être le reflet de différentes périodes d'occlusions distribuées sur une phase de geste glottale relativement fixe. Les données analysées par Maddieson sont constituées de segments situés à l'initiale de mot, comme c'est le cas pour notre étude. Il sera donc nécessaire d'élargir l'étude à des éjectives en position médiane de mot pour permettre de mesurer la durée d'occlusion et ainsi vérifier si le patron observé pour les pulmonaires peut être généralisé aux éjectives.

Nos derniers résultats montrent que les éjectives avaient une tendance à avoir un VOT plus long devant la voyelle basse [a], et un VOT plus court devant la voyelle haute [i], avec un VOT intermédiaire devant [e]. Les résultats sont donc différents de ce qui a été rapporté pour les occlusives pulmonaires, où le VOT est significativement plus long devant [i] que devant [a]. Le fait que l'on ne retrouve pas le même résultat pour les éjectives nous interroge sur les contraintes articulatoires ou aérodynamiques en jeu dans le contexte vocalique. Cependant, nous devons rappeler que cette étude s'est effectuée sur deux locuteurs uniquement. Les conclusions à tirer doivent donc être appréhendées avec prudence. Il est ainsi nécessaire d'élargir le champ d'investigation à plus de locuteurs, et plus de données pour tenter de mieux appréhender la nature des variations du VOT en lien avec le lieu d'articulation et la nature des voyelles qui suivent. Pour ce faire, des analyses aérodynamiques sont prévues, et deux autres langues, le tigrinya et le mehri, sont en cours d'analyse pour déterminer si les mêmes caractéristiques observées en maya sont aussi attestées dans des langues non apparentées.

Remerciements

Les auteurs remercient les deux locuteurs mayas pour leur disponibilité et leur aide lors de l'acquisition des données acoustiques. 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. (2012). *Praat: doing phonetics by computer [Computer program]*. http://www.praat.org/

BRICKER, V., YAH, E. P. & POʔOT, O. D. (1998). *A Dictionary of The Maya Language As Spoken in Hocabá, Yucatán*. Salt Lake City: University of Utah Press.

CATFORD, J. C. (1939). On the classification of stop consonants, *Le Maître Phonétique*, 65, 2-5.

CHANG, S. S., OHALA, J. J., HANSSON, G. & JAMES, B. (1999). Vowel-dependent VOT variation: An

experimental study. *Journal of the Acoustical Society of America* 105(2) pt. 2, 1400.

CHO, T & LADEFOGED, P. (1999). Variation and Universals in VOT: Evidence from 18 Languages. *Journal of Phonetics* 27, 207-229.

COLAZO-SIMON, A. (2007). *Les phénomènes glottaux en situation de contact linguistique : maya et espagnol du Yucatán, Mexique*, Paris: Université Paris III - Sorbonne Nouvelle, thèse.

HIGGINS, M. B., NETSELL, R. & SCHULTE, L. (1998). Vowel related differences in laryngeal articulatory and phonatory function. *Journal of Speech, Language, and Hearing Research* 41, 712-724.

KINGSTON, J. (1985). *The phonetics and phonology of the timing of oral and glottal events*. Berkeley, CA: UC Berkeley, thèse.

LADEFOGED, P. & MADDIESON, I. (1996). *Sounds of the world's languages*. Oxford: Blackwells.

LEWIS, M. P. (2009). *Ethnologue: Languages of the world*. (16ème édition). Dallas, Tex.: SIL International. Version en ligne : http://www.ethnologue.com/, accès : 2016.

LINDAU, M. (1984). Phonetic Differences in Glottalic Consonants. *Journal of Phonetics* 12, 147–155.

LISKER, L. & ABRAMSON, A.S. (1964). A cross-language study of voicing in initial stops: Acoustic measurements. *Wor*d 20, 384-422.

MADDIESON, I. (2001). Good timing: Place dependent VOT in ejective stops. *Proceedings of Eurospeech*. 823-826.

MADDIESON, I., SMITH, C. L., & BESSELL, N. (2001). Aspects of the phonetics of Tlingit. *Anthropological Linguistics*, 135-176.

WARNER, N. (1996). Acoustic Characteristics of Ejectives in Ingush. *Proceedings of the International Conference on Spoken Language Processing*, 1525–1528. Philadelphia, PA.

WEISMER, G. (1979). Sensitivity of Voice-Onset-Time (VOT) measures to certain segmental features in speech production. *Journal of Phonetics* 7, 197-204.

WRIGHT, R., HARGUS, S., & DAVIS, K. (2002). On the categorization of ejectives: Data from Witsuwit'en. *Journal of the International Phonetic Association* 32, 43-77.

WYSOCKI, T. (2004). *Acoustic analysis of Georgian stop consonants and clusters.* Chicago: University of Chicago, thèse.

Accommodation temporelle chez l'enfant dans une tâche de parole alternée

Céline Hidalgo[1, 2] Simone Falk[1, 3] Daniele Schön[2]
(1) Laboratoire Parole et Langage, UMR 7309, CNRS, Aix-en-Provence, France
(2) Institut de Neurosciences des Systèmes UMR 1106, INSERM, Marseille, France
(3) Institut de Philologie allemande, Ludwig-Maximilians-Universitaet, Munich, Allemagne

celine.hidalgo@univ.etu-amu.fr, Simone.Falk@germanistik.uni-muenchen.de, daniele.schon@univ-amu.fr

RESUME

L'accommodation temporelle entre deux interlocuteurs est un phénomène qui émerge lors d'une interaction et qui jouerait un rôle important dans la fluidité des échanges. Cette étude examine cette capacité temporelle chez l'enfant âgé de 5 à 6 ans grâce au développement d'une nouvelle tâche de dénomination en alternance avec un partenaire virtuel. Les variables temporelles analysées sont le tempo de l'alternance (*lent* versus *rapide*) et la rythmicité des mots échangés (*constante* versus *aléatoire*). Les enfants sont plus précis dans la condition de tempo rapide et plus réguliers lorsque la rythmicité des listes de mots est maintenue constante. Ces résultats montrent 1) que la dénomination en alternance est un paradigme permettant de mesurer les capacités d'accommodation temporelle des enfants et que 2) dès 5 ans, les enfants peuvent ajuster leur parole à celle d'un agent. Ces données constituent une base pour mesurer les capacités linguistiques d'accommodation temporelle chez des populations cliniques.

ABSTRACT

Children's temporal accommodation in an alternated naming task. Temporal accommodation between interlocutors is a phenomenon taking place in language interactions that may play an important role in improving communication. This study examines temporal accommodation skills of children aged from 5 to 6 years using a new paradigm. In a picture naming task in alternation with a virtual agent, we analyzed children's temporal accommodation according to two temporal parameters : speed of alternation - *fast* or *slow* - and the number of syllables of the words - *match* or *mismatch* - pronounced by the agent and children. Children were more accurate in the fast condition and more regular in the match condition. These results show that 1) naming in alternation is an effective task to measure temporal accommodation of children and that 2) children from age 5 adjust their speech to the timing of an agent. This study constitutes a baseline to assess language interaction abilities in clinical populations.

MOTS-CLES : accommodation interpersonnelle, enfants, prédiction temporelle, tour de parole, interaction, agent.

KEYWORDS : interpersonal accommodation, children, timing prediction, turn-taking, interaction, agent.

1 Introduction

La conversation est une situation co-construite par les interlocuteurs qui évolue au cours du temps. La dynamique de cette co-construction, est sous-tendue par un phénomène ténu mais néanmoins essentiel : la convergence. Des études ont en effet montré qu'au cours d'une interaction, les locuteurs ont une tendance à imiter certaines caractéristiques verbales- (phonétiques : Pardo, 2006 ; syntaxiques : Branigan et al., 2000)-et paraverbales (Krivokapić, 2013) de leur interlocuteur ou encore à s'aligner conceptuellement avec lui (Garrod & Pickering, 2004). Ces procédés nommés *convergence*, *alignement*, *accommodation*, *synchronisation*, permettraient d'une manière générale de favoriser une bonne compréhension réciproque. Plus spécifiquement, ces phénomènes appliqués au niveau temporel permettraient de fluidifier la dynamique conversationnelle. En effet, en mimant certaines caractéristiques temporelles de la parole de son interlocuteur, l'auditeur augmenterait sa capacité à prédire la fin des tours de parole. Ces prédictions temporelles lui permettraient ainsi d'anticiper la programmation de sa propre parole afin qu'elle soit produite sans trop de délai ni trop de chevauchement lors de son tour de parole.
Selon Wilson & Wilson (2005), ce mécanisme de projection lors de la prise des tours entre les interlocuteurs se réalise au niveau cérébral grâce au couplage de phase phase-locking ») entre la fréquence d'apparition des syllabes du locuteur et l'activité oscillatoire spontanée de ses neurones. Ainsi, le système moteur du locuteur et le système auditif de l'auditeur entreraient en couplage de phase à partir de la fréquence du rythme de parole (débit ou nombre de syllabes par seconde) qui est déterminé de manière dynamique (donc variable) au cours de l'interaction par les mouvements oscillatoires de la mandibule des interlocuteurs (Scott et al., 2009). Ce phénomène de « phase-locking » neuronal permet aux interlocuteurs de s'accommoder par un processus dit d' « entrainement » au rythme de la parole de l'autre (Peelle & Davis, 2012). La fréquence oscillatoire du débit de parole de l'autre étant alors simulée de manière endogène, l'auditeur peut ainsi facilement émettre des prédictions sur le moment où le tour de parole de son interlocuteur va prendre fin et sur le moment où son tour, en tant que locuteur, va devoir commencer ; la programmation temporelle de sa parole étant conditionnée par ces processus anticipatoires (Garrod & Pickering, 2015).
De récentes études utilisant la technique des mouvements oculaires (Casillas & Frank, 2013) ont montré que les enfants sont capables, dès trois ans, d'anticiper l'occurrence temporelle des tours de parole de deux interlocuteurs en train de dialoguer ; ce phénomène étant plus particulièrement marqué lors de la présence d'énoncés hautement prévisibles tels que les paires de questions-réponses. En production, les études s'intéressant au développement de cette compétence s'attachent majoritairement à évaluer la qualité de synchronisation de dyades mère-bébé à un niveau préverbal (Jaffe et al., 2001). Il semble cependant intéressant de coupler les recherches récentes en perception chez l'enfant d'âge verbal à des tâches en production, permettant de mesurer leurs capacités de convergence temporelle dans le cadre d'une interaction. Pour ce faire, nous avons créé une tâche dans laquelle un enfant dénomme des images en alternance avec un partenaire virtuel. Ce type de paradigme nous a permis de manipuler deux paramètres temporels : le tempo de l'alternance et la constance de la rythmicité des mots échangés. Ces deux facteurs nous permettent de mesurer la convergence de l'enfant vers une temporalité interactionnelle imposée.

2 Matériel et Méthode

2.1 Population

Nous avons recruté un groupe de 16 enfants, âgés de 5 et 6 ans (étendue = 52 mois à 71 mois ; moyenne d'âge = 65 mois ; écart-type= 5 mois), de langue maternelle française. Ce groupe était composé de 10 filles et 6 garçons. Les enfants porteurs de troubles visuels, de troubles du langage ou de troubles auditifs ont été exclus de l'étude. Ces enfants francophones étaient tous scolarisés dans une école du centre de Marseille en classe de moyenne et de grande section de maternelle. Nous avons reçu le consentement éclairé de tous les parents avant de commencer l'étude. Les enfants ont tous reçu un livre en remerciement de leur participation.

2.2 Stimuli

Pour cette tâche de dénomination, nous avons réalisé une sélection de 160 images comportant 80 mots monosyllabiques et 80 mots bisyllabiques issues de la banque de d'images BD2I (Cannard et al., 2006) en fonction de leur taux de dénomination pour un âge d'acquisition allant de 3 à 8 ans. La moyenne du taux de réussite de dénomination pour les mots monosyllabiques est de 93,9, écart-type= 5,3 et la moyenne du taux de réussite de dénomination pour les mots bisyllabiques est de 98,8, écart-type= 1,6. Le lexique représenté par ces images était composé de 17 catégories taxonomiques : aliment, animal, bijou, fourniture scolaire, habitation, instrument de musique, jouet, meuble, outil, partie du corps, paysage, personne, plante, ustensile de cuisine, véhicules, vêtements, autre. A partir des images sélectionnées, nous avons créé des paires d'images : la première image de la paire étant destinée au partenaire virtuel (PV) et la deuxième à l'enfant (figure 1). La composition de ces paires était différente en fonction des listes de mots créées. Dans les listes à rythmicité constante (condition *match*), les paires étaient composées de mots avec un nombre de syllabes identique (i.e. dans certaines listes toutes les paires étaient composées de mots monosyllabiques et dans d'autres, toutes les paires étaient composées de mots bisyllabiques). Dans les listes à rythmicité aléatoire (condition *mismatch*), les paires étaient composées de mots avec un nombre de syllabes aléatoirement identique ou différent (i.e. certaines paires étaient composées de mots monosyllabiques ou bisyllabiques et, à l'intérieur de la même liste, certaines paires étaient composées de mots avec un nombre de syllabes différent : par exemple, le PV dénommait un mot monosyllabique alors que l'enfant avait à dénommer un mot bisyllabique). Nous avons ainsi créé 8 listes de 10 paires d'images chacune selon cette procédure : 4 listes *match* et 4 listes *mismatch*. Pour éviter des effets d'amorçage sémantique, catégoriel ou phonétique, nous avons pris soin de ne jamais mettre deux mots de la même catégorie, commençant par le même phonème ou appartenant à la même catégorie dans une paire d'images ou dans deux paires successives.

Les images destinées au PV ont été dénommées à débit normal et enregistrées à partir d'une voix féminine, dans une chambre sourde à un échantillonnage de 44.1 kHz.
Afin de générer la perception d'une rythmicité dans l'échange, nous avons extrait, grâce à un algorithme développé par Cummins & Port (1998), tous les points acoustiques déterminés par notre système perceptif comme étant le début de chaque syllabe (« p-centers ») des mots enregistrés et nous les avons fait délivrés, via le logiciel Presentation (Neurobehavioral System) à intervalles réguliers : soit toutes les 3200 ms (condition *lente* du tempo de l'alternance), soit toutes les 2600 ms (condition *rapide*).

L'ordre de présentation des listes à l'intérieur des 2 modalités du facteur Vitesse d'alternance, ainsi que l'ordre de présentation des modalités du facteur Rythmicité lors du test, ont été contrebalancés. Toutes les listes de paires d'images étaient précédées d'un exemple de 3 échanges entre un enfant et le PV en « parfaite » alternance afin de simuler la situation expérimentale et de familiariser l'enfant avec la rythmicité et le tempo de la liste à venir. Ainsi, pour ces 3 paires de mots, nous avons placé le p-center de l'enfant virtuel à la moitié de l'intervalle séparant les 2 p-centers successifs du partenaire virtuel (figure 2) c'est-à-dire à 1600 ms pour la modalité *lente* et 1300 ms pour la modalité *rapide*.

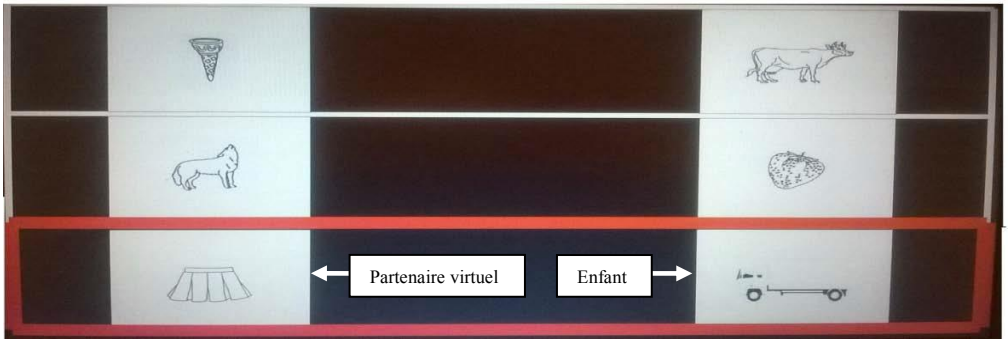


FIGURE 1- Capture d'écran de la tâche de dénomination. Les images de gauche sont dénommées par le partenaire virtuel et celles de droite par l'enfant. Le cadre rouge défile régulièrement sur l'écran au fur et à mesure de la succession des paires d'images.

Ces stimuli sonores ont été délivrés en utilisant une carte son de marque Creative via une enceinte de marque Sony. Les dénominations des images des enfants ont été enregistrées avec un micro-casque de marque Sennheiser. Les paires d'images ont été délivrées via un ordinateur portable de marque DELL, résolution 1366 X 768 (figure 1).

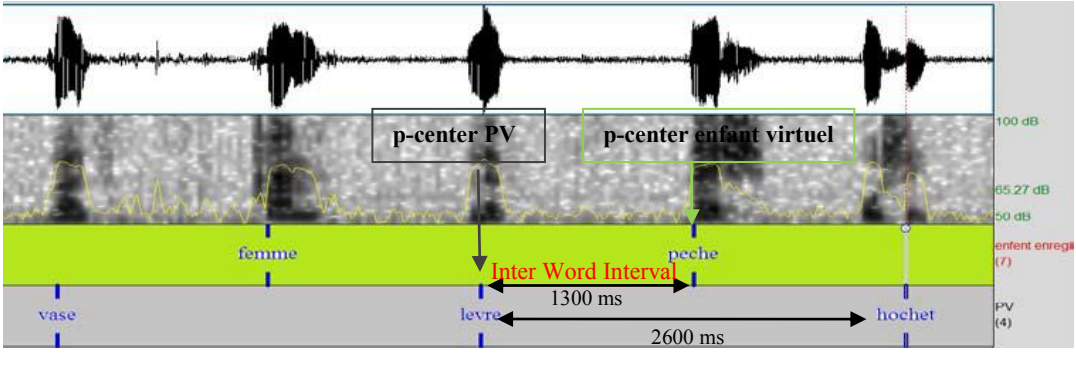


FIGURE 2- Textgrid (Praat) montrant le placement régulier des p-centers sur des paires de mots préenregistrées (PV+enfant virtuel) selon une rythmicité d'alternance « parfaite » délivrée avant chaque liste.

2.3 Procédure

Avant le test, une étape de familiarisation avec les images de chaque liste ainsi qu'avec le principe de l'alternance a été effectuée afin de s'assurer que l'enfant serait en mesure de réaliser la tâche convenablement, qu'il reconnaîtrait l'image de l'objet, que le lexique serait connu et que l'intelligibilité serait suffisante pour assurer les mesures et les analyses.

Lors du test, les images ont été présentées à distance et à intensité confortables en fonction de chaque enfant. La consigne était de regarder l'écran d'ordinateur et de dénommer la deuxième image de la paire à la suite de la dénomination de la première image par le PV. Il était explicitement demandé que la dénomination soit réalisée à la même vitesse que l'enfant qu'il venait d'entendre parler. L'enfant était encouragé entre chaque liste.
La succession des images n'a pas été interrompue même si l'enfant ne se souvenait plus du lexique représenté par l'image afin de ne pas rompre la rythmicité créée par l'alternance. Chaque enfant a été testé individuellement par le même expérimentateur dans une salle d'étude d'une école du centre de Marseille.

2.4 Mesures

Afin d'analyser si l'enfant s'améliorait dans la perception et la réalisation de la régularité alternée en fonction des conditions, nous avons mesuré l'espace séparant le p-center de chaque mot du PV et celui de l'enfant pour chaque paire d'images. Cet intervalle, appelé Inter-Word-Interval (IWI) (figure 2 en rouge) servait de base pour le calcul de l'accommodation. Les p-centers ont été extraits de manière semi-automatique (Cummins & Port, 1998), leur localisation inspectée et, si nécessaire, corrigée manuellement dans le logiciel PRAAT (Boersma, 2002). Les mots précédés par un « euh » d'hésitation, un article, un bredouillement (duplication du 1er phonème) ont été exclus des analyses. De même, nous avons éliminé les essais (151 sur 1280) lorsqu'un temps de latence trop long de la part de l'enfant a généré un chevauchement de parole avec le mot suivant du partenaire virtuel.

2.5 Analyses statistiques

A partir de notre mesure, *l'Inter-word-Interval (IWI)*, qui est l'intervalle de temps qui sépare le p-center du PV de celui de l'enfant, nous avons conduit deux types d'analyses. Nous avons tout d'abord réalisé, à l'aide du logiciel Statistica, une analyse de la variance à mesures répétées avec deux facteurs intra-sujets : Vitesse (*lent* versus *rapide*) et Rythmicité (*match* versus *mismatch*), le seuil de significativité étant fixé à α=.05. Nous avons ensuite conduit, à l'aide de la boîte à outils CircStat du logiciel Matlab, des statistiques circulaires descriptives. Ce type d'analyses, nous a permis de transposer les réponses de chaque enfant dans un espace vectoriel dont la longueur du cercle représente un cycle oscillatoire de 2600 ms en condition *rapide* et 3200 ms en condition *lente*. Le point 0 (ou 360°) représente le placement du p-center du PV et le point à l'antiphase du point 0, c'est à dire à 180°, représente le moment « attendu » du placement du p-center de l'enfant. Les points situés dans l'espace supérieur à ce point traduisent les réponses trop précoces et les points situés dans l'espace inférieur les réponses trop tardives par rapport à une parfaite régularité rythmique entre les productions des locuteurs. La moyenne de tous ces points a ensuite été représentée par un vecteur d'angle Thêta (exprimé en radians) et de longueur r. L'angle de ce vecteur permet de déterminer la précision des réponses de l'enfant par rapport au moment attendu ou encore le niveau d'*asynchronie* : plus l'angle est grand, plus les réponses de l'enfant sont éloignées du moment attendu. La longueur du vecteur permet de déterminer la consistance de

ses réponses, c'est à dire leur degré de variabilité : plus le vecteur est long, plus la *consistance des réponses* de l'enfant est importante. Nous avons tout d'abord réalisé le test d'hypothèses de Rayleigh afin de contrôler le caractère non aléatoire de la distribution des réponses autour du cercle pour chaque enfant, le seuil de significativité étant fixé à α=.05. La distribution relevée dans chaque condition ne révélant pas de distribution au niveau du hasard ($p<.05$), nous avons pu procéder à l'analyse de l'angle et de la longueur du vecteur. A partir de la moyenne des angles Thêta, obtenue en radians dans chaque condition et pour chaque enfant, nous avons transformé ces radians en millisecondes puis réalisé une analyse de la variance à mesures répétées avec les facteurs intra-sujets *Vitesse* et *Rythmicité*. A partir de la moyenne de la longueur des vecteurs obtenue dans chaque condition, nous avons réalisé une analyse de la variance avec les mêmes facteurs intra-sujets que pour les angles (*Vitesse* et *Rythmicité*).

3 Résultats

3.1 Prise en compte de la vitesse et influence sur la précision des réponses

La figure 3 montre un effet significatif de vitesse. Les enfants ont en effet placé leurs mots en

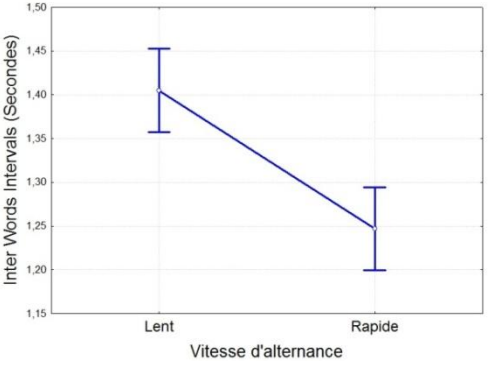


FIGURE 3 - Intervalles en secondes entre les p-center de l'enfant et ceux du partenaire virtuel en fonction de la vitesse d'alternance. Les barres verticales représentent les intervalles de confiance à 95%.

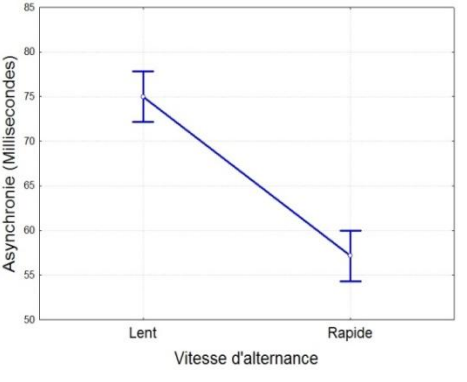


FIGURE 4 - Degré d'asynchronie en millisecondes (mesure d'imprécision) du placement du p-center de l'enfant par rapport au moment attendu en fonction des conditions de vitesse de l'alternance. Les barres verticales représentent les intervalles de confiance à 95%.

fonction du tempo d'alternance imposé, conservant ainsi une alternance plus rapide dans la condition *rapide* par rapport à la condition *lent* $F(1, 15) = 24.988$, $p<.000$. Les mesures effectuées à partir des statistiques circulaires sur la *grandeur de l'angle* révèlent également un effet principal de vitesse sur l'*asynchronie*. Les enfants sont moins précis dans la condition *lent* que dans la condition *rapide :* $F(1, 15)=89.710$, $p<.000$ (figure 4).

3.2. Consistance des réponses : effet de vitesse et influence de la régularité de rythmicité

Les mesures effectuées à partir de la *longueur du vecteur* révèlent que les enfants répondent de manière plus *consistante* dans la condition où la vitesse d'alternance est *rapide* comparé à la condition *lente* $F(1, 15)=6.852$, $p=.019$ et que leurs réponses sont plus consistantes pour les listes à rythmicité *constante* plutôt qu'*aléatoire* $F(1,15)=4.093$, $p=.061$ (effet marginal) (figure 5).

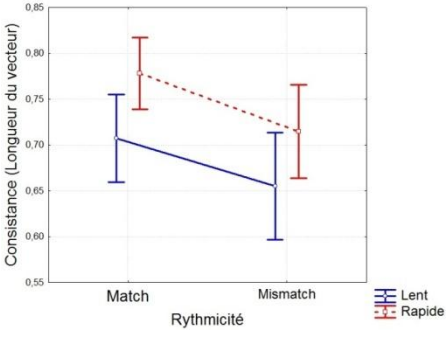


FIGURE 5 - Degré de consistance du placement des p-centers en fonction de la vitesse d'alternance et de la rythmicité des listes de mots. Les barres verticales représentent les intervalles de confiance à 95%.

4 Discussion et Conclusion

4.1 Effet de la vitesse de dénomination sur le système perceptivo-moteur

Les résultats obtenus à partir de ce nouveau paradigme montrent non seulement que les enfants de 5-6 ans sont capables d'adapter leur vitesse de dénomination à celle imposée par un partenaire virtuel mais également que notre tâche est pertinente pour tester l'accommodation temporelle chez des enfants de cette tranche d'âge. En outre, le fait que les enfants répondent au-delà du hasard dans toutes les conditions et qu'ils soient systématiquement en avance par rapport au moment de dénomination attendu nous prouve que les données récoltées ne sont pas du fait de la latence de dénomination mais bien de leur convergence temporelle. De plus, il est intéressant de noter que les enfants sont plus proches du temps attendu en condition *rapide* qu'en condition *lente*. Cette précision temporelle de la dénomination de l'enfant par rapport au PV est le facteur qui conditionne la réalisation d'une régularité parfaite de l'alternance. Cette précision temporelle accrue en condition de vitesse rapide génère une alternance des mots (PV+enfant) toutes les 1300 ms contre 1600 ms en vitesse lente. Dans les quelques données connues sur le développement du timing du tour de parole chez l'enfant, les études ont pu montrer qu'à 3 ans, les enfants qui échangent entre eux présentent en moyenne des transitions de tour de parole situées autour de 1500 ms mais qu'à 4 ans, ils se rapprochent du timing des adultes situé entre 20 et 1000 ms (Casillas, 2014). On peut alors penser que la vitesse proposée en condition *rapide* est plus proche d'une vitesse « naturelle » d'échange d'un enfant de 5 ans que la vitesse lente. D'autre part, l'augmentation de la consistance des réponses à cette vitesse de dénomination, suggère que dans la condition *rapide*, le système sensori-moteur de l'enfant arrive mieux à se synchroniser sur le stimulus de parole. Les études portant sur la synchronisation sensorimotrice telles que celles utilisant la technique de « tapping », montrent que la synchronisation avec un métronome devient

difficile dès lors qu'on se rapproche de 1800 ms (MacDorman, 1962, cité dans Repp & Doggett, 2007). La condition lente générant une succession des p-centers toutes les 1600 ms, elle place le système perceptivo-moteur des enfants dans une condition plus proche de cette limite de perception de l'isochronie que la condition *rapide* (1300 ms).

4.2 Effet de la rythmicité sur le système perceptivo-moteur

En faisant varier le nombre de syllabes des mots échangés entre le partenaire virtuel et l'enfant, nous souhaitions analyser si la perturbation de la rythmicité de l'alternance pouvait avoir une influence sur la régularité de dénomination des enfants ; ceci nous renseignant non seulement sur la capacité des enfants de 5-6 ans à percevoir une rythmicité induite dans une alternance de parole mais également à la reproduire. Les données issues des statistiques circulaires mettent en exergue la sensibilité des enfants à cette variation confirmant ainsi notre hypothèse. La plus grande consistance des réponses retrouvées dans la condition où la rythmicité des paires de mots est demeurée constante (condition *match*) montre que les enfants sont sensibles à une rythmicité induite dans un rythme de parole mais aussi que celle-ci a une conséquence sur la programmation de leur parole : les enfants sont plus réguliers lorsque la rythmicité des paroles échangées entre les locuteurs est régulière comparée à irrégulière. Ainsi, si l'on appréhende ce résultat selon les mécanismes de la dynamique oscillatoire neuronale, on peut penser qu'une régularité dans la rythmicité des paires de mots échangées permet plus facilement l'émergence du phénomène d' « *entrainment* » c'est-à dire du couplage de phase entre le stimulus de parole et les populations neuronales du système auditif (Large & Jones, 1999) ; l'émergence plus évidente d'une métrique dans la condition de rythmicité constante étant probablement à la base de ce phénomène facilitateur (Nozaradan et al., 2012). En outre, il est également possible que cette condition ait favorisé un couplage de phase entre les systèmes auditifs et moteurs (Large et al., 2015), augmentant la génération de prédictions temporelles sur les p-centers à venir ainsi que la régularité de la distribution des dénominations.

4.3 Dénomination alternée - un paradigme utile pour la recherche sur les capacités d'interaction chez l'enfant?

Si l'on transpose les effets obtenus avec ce nouveau paradigme dans le cadre d'une interaction naturelle, cela signifierait que l'enfant parvient, dans certaines conditions de tempo conversationnel, à prendre son tour de parole au moment *le plus opportun* dans la conversation c'est-à-dire en ne chevauchant pas la parole de son interlocuteur, et en ne laissant pas trop délai entre le tour de parole de celui-ci et le sien. En outre, les effets observés suite à la manipulation de la rythmicité de l'alternance montrent que la régularité de la parole de l'autre influence la capacité de l'enfant à s'accommoder et à demeurer régulier dans la programmation de sa parole.

Certaines conditions temporelles de l'alternance ont donc favorisé chez l'enfant le phénomène d'anticipation utilisé dans la gestion des tours de parole (Pickering & Garrod, 2015). Ceci nous permet de penser que grâce à ce type de paradigme et à la manipulation de facteurs temporels tels que la rythmicité de parole, des compétences temporelles sous jacentes à la fluidité de la dynamique interactionnelle pourront désormais être évaluées chez des populations cliniques d'enfants souffrant de troubles du langage. En outre, l'effet de la manipulation de la rythmicité sur la régularité de la production de parole nous conforte dans le fait que des techniques thérapeutiques, telles que la musique, visant au développement du traitement temporel auditif, devraient plus systématiquement être proposées dans la prise en charge des troubles du langage chez l'enfant.

Remerciements

Cette étude a été réalisée grâce au soutien financier du Brain and Language Research Institute (BLRI, ANR-11-LABX-0036), Laboratoire Parole et Langage - CNRS & Université d'Aix-Marseille.

Références

BOERSMA, P. (2002). Praat, a system for doing phonetics by computer. *Glot international* 5, 341-345.

BRANIGAN, H. P., PICKERING, M. J., CLELAND, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition* 75, B13–B25.

CANNARD, C., BLAYE, A., SCHEUNER, N., & BONTHOUX, F. (2005). Picture naming in 3-to 8-year-old French children : Methodological considerations for name agreement. *Behavior Research Methods* 37, 417-425.

CASILLAS, M. (2014). Pragmatic development in first language acquisition. Amsterdam/Philadelphia : Matthews, D.

CASILLAS, M., FRANK, M. C. (2013). The development of predictive processes in children's discourse understanding. In CogSci 2013 *The 35th annual meeting of the Cognitive Science Society*, 299-304.

CUMMINS, F., PORT, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics* 26, 145-171.

GARROD, S., PICKERING, M. J. (2015). The use of content and timing to predict turn transitions. *Frontiers in Psychology* 6, 1–12.

JAFFE, J., BEEBE, B., FELDSTEIN, S., CROWN, C. L., JASNOW, M. D., ROCHAT, P., STERN, D. N. (2001). *Rhythms of dialogue in infancy : Coordinated timing in development*. Temple University : W. F., Overton.

KRIVOKAPIĆ, J. (2013). Rhythm and convergence between speakers of American and Indian English. *Laboratory Phonology* 4, 39–65.

LARGE, E. W., HERRERA, J. A.,VELASCO, M. J. (2015). Neural Networks for Beat Perception in Musical Rhythm. *Frontiers in Systems Neuroscience* 9.

LARGE, E. W., JONES, M. R. (1999). The dynamics of attending : How people track time-varying events. *Psychological Review* 106, 119–159.

NOZARADAN, S., PERETZ, I., MOURAUX, A. (2012). Selective neuronal entrainment to the beat and meter embedded in a musical rhythm. *The Journal of Neuroscience* 32, 17572–81.

PARDO, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119, 2382-2393.

PEELLE, J. E., DAVIS, M. H. (2012). Neural Oscillations Carry Speech Rhythm through to Comprehension. *Frontiers in Psychology* 3, 1-17.

REPP, B. H., DOGGETT, R. (2007). Tapping to a very slow beat : A comparison of musicians and nonmusicians. *Music Perception* 24, 367–376.

SCOTT, S. K., MCGETTIGAN, C., EISNER, F. (2009). A little more conversation, a little less action-candidate roles for the motor cortex in speech perception. *Nature Reviews Neuroscience* 10, 295-302.

WILSON, M.,WILSON, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review* 12, 957–68.

Accès lexical et reconnaissance du voisement en voix chuchotée

Yohann Meynadier, Sophie Dufour
Aix Marseille Université, CNRS, LPL UMR 7309, 13100, Aix-en-Provence, France
yohann.meynadier@lpl-aix.fr, sophie.dufour@lpl-aix.fr

RÉSUMÉ

La reconnaissance du trait de voisement de consonnes obstruantes chuchotées en français a été examinée via un paradigme d'amorçage sémantique auditif-visuel. Un effet d'amorçage d'amplitude similaire à celui mesuré en voix modale a été observé uniquement lorsque l'obstruante du mot amorce chuchoté est sourde (*dessert*-CHOCOLAT). Aucun effet d'amorçage n'a été noté quand l'obstruante du mot amorce est voisée (*désert*) que ce soit sur le mot cible SABLE associé sémantique de *désert* ou sur le mot cible CHOCOLAT associé sémantique de *dessert*. Ainsi, même si certaines travaux ont mis en évidence qu'en voix chuchotée les consonnes obstruantes voisées maintiennent des traces phonétiques de leur identité sous-jacente, notre étude montre que ces consonnes sont ambigües pour l'auditeur et que leur reconnaissance n'est pas immédiate.

ABSTRACT

Lexical acces and recognition of voicing in whisper

The recognition of the voicing feature of whispered obstruant consonants was examined in a cross modal semantic priming paradigm. A priming effect of similar magnitude to that observed in modal voice was found only when the whispered prime includes a voiceless obstruant (e.g. *dessert* primes CHOCOLAT). No priming effect was found when the whispered prime includes a voiced obstruant (*désert*) neither on the target word SABLE semantically related to *desert*, nor on the target word CHOCOLAT semantically related to *dessert*. Hence, although a few studies have shown that whispered voiced obstruent consonants retain phonetic traces of their underlying identity, our study shows that these consonants are ambiguous for the listeners, and that their recognition is not immediate.

MOTS-CLÉS : perception du voisement, dévoisement, voix chuchotée, accès lexical
KEYWORDS: voicing perception, devoicing, whisper, lexical access

1 Introduction

Cette étude porte sur la perception du trait de voisement en parole chuchotée en français. Une consonne voisée en voix chuchotée est produite sans vibration des cordes vocales, constituant en voix modale la propriété principale du trait [+voisé]. On peut donc penser qu'une consonne voisée chuchotée est ambiguë avec la consonne sourde de mêmes lieu et mode d'articulation.

Or, des études de production en français montrent qu'en voix chuchotée les consonnes obstruantes voisées maintiendraient des traces phonétiques de leur identité sous-jacente : durée consonantique (Vercherand, 2010 ; Meynadier, Gaydina, 2013), durée de la voyelle préconsonantique (Meynadier, Gaydina, 2013), pression intraorale (Meynadier, Gaydina, 2013 ; Garnier et al., 2014 ; Meynadier 2015), pression de contact (Garnier et al., 2004) ou aire glottique (Malécot, Peebles, 1965 ; Crevier-

Buchman et al., 2009 ; Meynadier, 2015). Ces observations rejoignent les faits d'assimilation de voisement en français. Snoeren et al. (2008) et Hallé et al. (2012) relèvent en effet que les obstruantes sourdes et voisées assimilées, et les voyelles pré-consonantiques, conservent la durée caractéristique liée à leur voisement sous-jacent.

Les quelques études sur la perception du contraste de voisement en voix chuchotée en français sont peu comparables et montrent des résultats contradictoires. Vercherand (2010) observe un taux de reconnaissance des obstruantes chuchotées de plus de 80 %, dans un test d'identification de mots à choix multiple (2-4 choix). Dans une tâche d'identification libre de syllabes CV, Fux (2012) relève, selon le lieu d'articulation de l'obstruante, des taux variant de 15,6 à 46,7 % pour les voisées et de 80 à 97,8 % pour les sourdes. Inversement, dans un test d'identification à 2 choix forcés sur des non-mots en paire minimale, Meynadier et al. (2013) trouvent de façon inattendue que les obstruantes sourdes ne sont pas reconnues (56,7 %, choix au hasard), alors que les voisées très bien (89,2 %). Par contre, ils observent aussi que des modifications de la durée consonantique et/ou des voyelles précédentes affectent le taux de reconnaissance des obstruantes chuchotées.

Nous proposons ici de reprendre cette question via un paradigme expérimental d'amorçage sémantique auditif-visuel, fréquemment utilisé pour tester la reconnaissance de mots ambigus (Tabossi, 1996). Avec ce paradigme, Snoeren et al. (2008) ont pu montré que les auditeurs français peuvent de reconnaître le trait de voisement d'une obstruante sourde phonétiquement voisée par une assimilation totale. Cette reconnaissance du voisement sous-jacent s'appuierait sur un traitement des détails phonétiques présents dans la consonne pour accéder à la forme abstraite du mot.

Dans notre étude, des paires minimales basées sur un contraste de voisement (e.g. *dé***s***ert/de***SS***ert*) ont été présentées dans un paradigme d'amorçage sémantique auditif-visuel. Les participants entendaient d'abord un mot amorce en voix chuchotée ; puis immédiatement après, ils recevaient une cible présentée visuellement sur laquelle ils devaient réaliser une tâche de décision lexicale, consistant à décider le plus rapidement possible si une séquence de lettres constitue ou non un mot du lexique du français. Dans les conditions critiques, la cible (e.g. SABLE) pouvait être précédée du membre de la paire minimale qui lui était sémantiquement reliée (e.g. *désert*) ou de l'autre membre de la paire minimale (e.g. *dessert*). Les temps de réponse dans ces conditions ont été comparés à ceux d'une condition contrôle où l'amorce et la cible n'avaient aucun lien sémantique entre elles (e.g. *jumelle*–SABLE).

Dans l'hypothèse où le trait de voisement [+voisé] est automatiquement récupéré en parole chuchotée, on dervrait observer dans la condition d'appariement de voisement entre la cible et l'amorce (e.g. *désert*–SABLE) que l'amorce *désert* facilite le traitement subséquent de la cible SABLE, en comparaison à la condition contrôle (e.g. *jumelle*–SABLE). Par contre, dans la condition de non appariement (e.g. *désert*–CHOCOLAT), l'amorce chuchotée *désert* ne devrait pas faciliter le traitement de l'associé sémantique CHOCOLAT de l'autre membre de la paire minimale (i.e. *dessert*). Similairement, si le trait [-voisé] est automatiquement reconnu, *dessert* amorcerait CHOCOLAT, mais pas SABLE. Cependant, si en parole chuchotée la consonne voisée n'est pas reconnue et est perçue comme sourde, l'amorce *désert* ne faciliterait par la reconnaissance de la cible SABLE ; mais également, dans la condition de non appariement de voisement entre amorce et cible (e.g. *désert*–CHOCOLAT), on devrait observer que *désert* peut amorcer CHOCOLAT, mot cible associé sémantiquement au membre à consonne sourde de la même paire minimale (i.e. *dessert*).

2 Méthode

Nous présentons ici la sélection des couples de mots amorce et cible, leur validation pour le paradigme d'amorçage sémantique en voix modale et les tests expérimentaux en voix chuchotée.

2.1 Matériel linguistique

L'élaboration des couples de mots amorce-cible a nécessité un pré-test d'association sémantique libre pour les mots amorces. Les amorces ont été sélectionnées parmi 173 mots (dont un homophone non homographique) du lexique français faisant partie d'une paire minimale opposant en position intervocalique une obstruante voisée, parmi /b d g v z ʒ/, à son pendant sourd /p t k f s ʃ/ (e.g. *cousin/coussin*, *envie/amphi*). Cette liste de mots amorces a été soumise, dans cinq ordres aléatoires différents, à 38 sujets (8 hommes, 30 femmes, de 20 à 47 ans). Les sujets devaient fournir par écrit et le plus rapidement possible le premier mot leur venant à l'esprit à la lecture de chaque mot amorce. Seules ont été retenues 20 paires minimales de mots amorces dont l'associé sémantique le plus fréquent (i.e. le mot cible) de chaque membre de la paire a été donné par au moins 20% des sujets (e.g. *désert*–SABLE *vs dessert*–CHOCOLAT). Les taux d'association des cibles reliées à une amorce avec consonne voisée (abrégé en A*v*C*v*, comme *dé**s**ert*–SABLE) et à une amorce avec consonne sourde (abrégé en A*s*C*s*, comme *de**ss**ert*–CHOCOLAT) ne différaient pas significativement (Table 1, %asso).

	%asso	Nph	Ngr	Nsy	PU	FreCum	Durée mot		Durée obstruante	
							modal	chuch.	modal	chuch.
Amorce voisée (A*v*)		4,35		2,20	4,95	21,45	554	703	93	88
Amorce sourde (A*s*)		4,35		2,20	4,80	13,90	541	698	144	156
Amorce contrôle		4,35		2,20	4,85	15,95				
Cible voisée (C*v*)	41,20		5,50	1,75		44,05				
Cible sourde (C*s*)	40,30		5,90	1,80		105,79				

TABLE 1 : Caractéristiques des amorces et des cibles. Taux d'association sémantique (%asso) ; nombre de phonèmes (Nphon) ; nombre de lettres (Ngraph) ; nombre de syllabes (Nsyll) ; point d'unicité (PU) ; fréquence cumulée (FreCum) ; durée moyenne (ms) des mots et des obstruantes en opposition distinctive

Ainsi, 40 couples amorce-cible, dont 20 A*v* et 20 A*s*, ont été élaborés. 20 mots amorces servant de contrôle, aussi bien dans la condition A*v*C*v* (e.g. *jumelle*–SABLE) que dans la condition A*s*C*s* (e.g. *jumelle*–CHOCOLAT), appariés aux amorces testées A*v* et A*s* en nombre de phonèmes, en nombre de syllabes, en point d'unicité phonologique et en fréquence, ont été sélectionnés. Les caractéristiques des amorces et des cibles sont présentées dans la Table 1. A noter que les mots cibles voisés (C*v*) et sourds (C*s*) différaient en terme de fréquence.

6 expériences, détaillées dans la Table 2, ont été construites : (i) 4 testaient les conditions avec appariement de voisement entre amorce et cible : A*v*C*v* et A*s*C*s* en voix modale et en voix chuchotée ; (ii) les 2 autres testaient les conditions de non appariement de voisement entre amorce et cible en voix chuchotée : A*v*C*s* et A*s*C*v*. Pour chacune des 4 conditions (A*v*C*v*, A*s*C*s*, A*v*C*s* et A*s*C*v*), 2 listes ont été crées afin que chaque mot cible (SABLE ou CHOCOLAT) soit précédé des deux types d'amorce : contrôle (*jumelle*) *vs* A*v* (*désert*) ou A*s* (*dessert*), et qu'un même participant ne voit pas deux fois le même mot cible. Chaque liste expérimentale incluait ainsi 10 mots cibles précédés d'une amorce A*v* ou As et 10 mots cibles d'une amorce contrôle. De façon à réduire la proportion d'essais reliés à 20%, les listes expérimentales incluaient également 30 couples de mots non reliés (*légume*–GRIFFE) servant ainsi de remplisseurs. Enfin, pour les besoins de la tâche, 50 non-mots ont été créés en changeant une lettre de mots existants et non utilisés dans le matériel expérimental (e.g. VALADE, issu de 'salade'). Parmi les 50 non-mots, 10 étaient précédés d'une amorce « pseudo » reliée (*laitue*–VALADE*)* et les 40 autres d'une amorce sémantiquement non reliée (*bouteille*–FUDUR,

issu de 'futur'). Les couples amorce-cible de non-mots « pseudo » reliés ont été obtenus à partir des normes d'associations verbales de Ferrand et Alario (1998).

Expérience	Condition	**A**morce audio *(ex.)*	**C**ible visuelle *(ex.)*
n° 1a et 2a	appariée **A*v*C*v***	consonne **v**oisée *(déSert)*	associé **v**oisé *(sable)*
n° 1b et 2b	appariée **A*s*C*s***	consonne **s**ourde *(deSSert)*	associé **s**ourd *(chocolat)*
n° 3a	non appariée **A*v*C*s***	consonne **v**oisée *(déSert)*	associé **s**ourd *(chocolat)*
n° 3b	non appariée **A*s*C*v***	consonne **s**ourde *(deSert)*	associé **v**oisé *(sable)*

TABLE 2 : Conditions expérimentales d'appariement selon le voisement entre amorce et cible

Les amorces ont été enregistrées en chambre sourde et ont été lues trois fois en voix modale et en voix chuchotée par un locuteur masculin (le 1er auteur). Le signal acoustique a été enregistré avec EVA2® (Ghio et Teston, 2004). L'item le plus intelligible et sans bruits de bouche a été retenu. Segmentés manuellement sous Phonedit® [http://www.lpl-aix.fr/~lpldev/phonedit], le début et la fin du mot ont été identifiés au début des premières et à la fin des dernières modulations visibles sur l'oscillogramme. Le début des occlusives sourdes initiales de mot étant acoustiquement invisible, la durée de leur occlusion (avant la barre d'explosion) a été fixée à 70 ms pour les sourdes en voix modale, et en voix chuchotée à 85 ms pour les sourdes et 65 ms pour les voisées. Ces valeurs correspondent aux durées moyennes d'occlusion mesurées par Meynadier et Gaydina (2013) sur le même locuteur. Les durées moyennes des amorces auditives A*v* et A*s* en voix modale et chuchotée sont fournies dans la Table 1. L'intensité de chaque stimulus audio a été normalisée à l'intensité moyenne de l'ensemble des stimuli dans chaque modalité vocale.
L'absence de vibration des cordes vocales en voix chuchotée a été confirmée par l'absence totale de périodicité de l'onde acoustique dans les mots amorce. La durée plus courte des obstruantes voisées comparée aux sourdes (Tableau 1) observée en voix chuchotée comme en voix modale valide que les paires minimales chuchotées utilisées comme stimuli présentent bien le contraste de voisement attendu malgré l'absence de vibration laryngée (cf. Introduction).

2.2 Expériences 1a et 1b : amorçage en voix modale

Les expériences en voix modale avaient pour but de valider les couples amorce-cible sémantiquement reliés, correspondant aux seules conditions appariées en voisement A*v*C*v* et A*s*C*s*.

Les sujets ont passé ces tests dans une pièce calme. L'amorce auditive était écoutée dans un casque Sennheiser HD415 et le volume de sortie de l'ordinateur portable toujours réglé à un niveau sonore fixe, maintenant constante entre sujets la différence d'intensité entre voix chuchotée et voix modale. La cible était présentée immédiatement à la fin de la présentation de l'amorce et les participants avaient pour tâche de décider le plus rapidement et le plus précisément possible si la cible constituait un mot ou un non-mot. Les réponses ont été enregistrées à l'aide d'un boitier à deux boutons-poussoirs ; la réponse « mot » étant placée du côté de la main dominante du sujet. Le test de perception a été élaboré avec le logiciel Perceval® (André et al., 2003). Un entrainement sans feedback aux réponses sur 10 items non utilisés ultérieurement précédait le test d'une durée d'environ 15 mn. Les participants bénévoles ont signé une fiche de consentement éclairé.

24 sujets ont passé l'Expérience 1a et 24 autres l'Expérience 1b (Table 2). Aucun n'avait pris par au pré-test d'association libre entre amorce et cible. Ces 16 hommes et 32 femmes, dont 5 gauchers et 43 droitiers, de 18 à 45 ans ($\mu = 27,6$; $\sigma = 7,3$), étaient des étudiants de langue maternelle française sans trouble du langage, de l'audition ou visuel non corrigé.

2.3 Expérience 2a, 2b, 3a et 3b : amorçage en voix chuchotée

Les 4 expériences (2a, 2b, 3a et 3b, Table 2) en voix chuchotée étaient destinées à tester la capacité de reconnaissance du trait de voisement en l'absence de vibration des cordes vocales.

Ont participés à ces tests, 96 sujets de même profil que ceux des deux expériences en voix modale : étudiants, 18-45 ans (μ = 27,9 ; σ = 7,6), français maternel, sans troubles. Aucun n'avaient participés ni au pré-test d'association ni à l'Expérience 1a ou 1b. Ces tests ont été passés avec le même matériel technique et la même procédure que les Expériences 1a et 1b. Seuls les mots d'amorce auditive différaient, étant ici produits en voix chuchotée. Les 4 groupes de 24 participants ont été répartis dans chacune des 4 expériences correspondant à chaque condition expérimentale.

3 Résultats

Les analyses statistiques ont été effectuées sur les temps de réaction des réponses correctes. Des taux de réponse correcte (« mot » *vs* « non-mot ») supérieurs à 90 % pour chaque test montrent que la tâche a bien été réalisée par les sujets. Les temps de réaction ont été mesurés à partir du début de l'affichage de la cible. Les réponses ayant un temps de réaction supérieur à 1300 ms ont été exclues. Les effets d'amorçage ont été évalués au moyen de t-tests par sujets (t_s) et par items (t_i). Le seuil de significativité statistique de p est fixé à .05.

3.1 Effet d'amorçage en voix modale : Expériences 1a et 1b

Les résultats de l'Expérience 1a et 1b sont présentés dans la Table 3. 5,63 % des réponses ayant un temps de réaction trop long dans la condition A*v*C*v* et 4,38 % dans la condition A*s*C*s* ont été exclues de l'analyse.

Expé.	Condition	Amorce contrôle	Amorce reliée sémantiquement	Différence (contrôle – relié)
n° 1a	A*v*C*v* (*désert*-SABLE)	680 (71) *2 %*	648 (71) *2 %*	+31 *
n° 1b	A*s*C*s* (*dessert*-CHOCOLAT)	682 (67) *2 %*	653 (66) *2 %*	+29 *

TABLE 3 : Temps de réaction moyens en ms (écart type) et pourcentage moyen d'erreurs par sujets en voix modale. Ecart type entre parenthèses ; taux d'erreurs en italique ; différence statistique : * significative ; *ns* non significative

Que l'amorce comporte une consonne voisée (A*v*C*v*) ou sourde (A*s*C*s*), le temps de décision lexicale sur le mot cible est similaire (648-653 ms). Mais surtout, les analyses révèlent un effet d'amorçage significatif aussi bien pour la condition A*v*C*v* [$t_s(23) = 2.88$, $p < .05$; $t_i(19) = 3.18$, $p < .05$] que pour la condition A*s*C*s* [$t_s(23) = 2.72$, $p < .05$; $t_i(19) = 2.53$, $p < .05$]. Le mot cible est plus rapidement reconnu s'il est précédé de son amorce sémantiquement reliée que d'une amorce non reliée (contrôle). Ainsi, *désert* facilite la décision lexicale sur le mot SABLE, de même que *dessert*accélère la décision sur CHOCOLAT. Cet effet d'amorçage a une amplitude similaire dans les deux cas, se situant autour de 30 ms.

L'amorçage sémantique en voix modale est donc bien observé. Ces résultats valident les couples de mots amorce-cible utilisés pour les tests de reconnaissance du trait de voisement en voix chuchotée.

3.2 Effet d'amorçage en voix chuchotée : Expériences 2a, 2b, 3a et 3b

Les résultats ont été analysés selon les mêmes critères que dans les Expériences 1a et 1b. Les temps de réaction moyens des bonnes réponses en voix chuchotée sont présentés dans la Table 4. Le pourcentage de données rejetées est de 4,79 % pour la condition A*v*C*v*, 1,46 % pour A*s*C*s*, 2,08 % pour A*v*C*s*, 5,21 % pour A*s*C*v*.

Expé.	Condition	Amorce contrôle	Amorce reliée sémantiquement	Différence (contrôle – relié)
n° 2a	A*v*C*v* (*désert*-SABLE)	715 (97) *2 %*	718 (111) *1 %*	-3 *ns*
n° 2b	A*s*C*s* (*dessert*-CHOCOLAT)	694 (85) *4 %*	666 (81) *4 %*	+28 *
n° 3a	A*v*C*s* (*désert*-CHOCOLAT)	673 (97) *2 %*	696 (123) *2 %*	-23 *ns*
n° 3b	A*s*C*v* (*dessert*-SABLE)	723 (95) *2 %*	722 (118) *3 %*	-1 *ns*

TABLE 4 : Temps de réaction moyens en ms (écart type) et pourcentage moyen d'erreurs par sujets en voix chuchotée

Les analyses montrent un effet d'amorçage significatif uniquement dans la condition appariée A*s*C*s* [$t_s(23) = 2.43$, $p < .05$; $t_i(19) = 2.42$, $p < .05$]. Comme en voix modale, entendre *dessert* en voix chuchotée accélère la décision lexicale sur la cible visuelle CHOCOLAT. L'amplitude de l'effet facilitateur est proche de 30 ms, ce qui est similaire à celui obtenu en voix modale.

Dans la condition appariée A*v*C*v*, aucun effet d'amorçage n'est observé [$t_s(23) = 0.27$, $p > .20$; $t_i(19) = 0.34$, $p > .20$]. Contrairement à la voix modale, quand l'amorce chuchotée comporte une consonne voisée (*désert*) la décision sur la cible reliée sémantiquement (SABLE) n'est pas accélérée.

Également, dans les conditions non appariées A*v*C*s* [$t_s(23) = 1.58$, $p = 0.13$; $t_i(19) = 1.37$, $p = 0.18$] et A*s*C*v* [$t_s(23) = 0.02$, $p > .20$; $t_i(19) = 0.16$, $p > .20$], aucun effet d amorçage n'est observé. Ainsi de la même manière que *dessert* chuchoté n'influence pas la vitesse de décision sur SABLE (associé sémantique de *désert*), *désert* chuchoté n'a eu aucun impact sur le temps de traitement du mot cible CHOCOLAT (associé sémantique de *dessert*).

3.3 Effet de la modalité vocale : voix modale *vs* chuchotée

Des analyses additionnelles comparant voix modale et voix chuchotée ont été conduites de façon à examiner l'impact de la modalité vocale sur les temps de réaction. Deux analyses de variance, l'une sur les amorces à obstruante voisée (A*v*C*v*) et l'autre sur les amorces à obstruante sourde (A*s*C*s*) ont été conduites avec la modalité vocale (modale *vs* chuchotée) et le type d'amorce (reliée *vs* contrôle, c'est-à-dire non reliée) comme facteurs (Figure 1).

Les résultats obtenus pour les amorces voisées (A*v*C*v*) montrent un effet significatif de la modalité vocale par sujets et par items [$F_s(1,46) = 4.49$, $p < .05$; $F_i(1,19) = 40$, $p < .0001$]. Les temps de réaction sont en moyenne plus longs en voix chuchotée qu'en voix modale. De façon critique, l'interaction entre les facteurs modalité et type d'amorce est significative dans les analyses par sujets

[$F_s(1,46) = 5.43$, $p < .03$] et approche la significativité dans les analyses par items [$F_i(1,19) = 3.72$, $p = .07$]. Comme nous l'avons vu au préalable, cette interaction résulte d'un effet d'amorçage significatif uniquement pour la voix modale, mais aussi d'un effet plus important de la modalité vocale pour les amorces reliées [$F_s(1,46) = 6.63$, $p < .02$; $F_i(1,19) = 29.85$, $p < .0001$], que pour les amorces contrôle [$F_s(1,46) = 2.09$, $p = .15$; $F_i(1,19) = 6.28$, $p < .03$].

Les résultats relatifs aux amorces sourdes (A*s*C*s*) montrent un effet de la modalité vocale proche de la significativité uniquement dans les analyses par items [$F_s(1,46) = 0.33$, $p > .20$; $F_i(1,19) = 4.24$, $p = .053$]. Aucune interaction entre les facteurs modalité et type d'amorce n'est observée [$F_s(1,46) = 0.006$, $p > .20$; $F_i(1,19) = 0.07$, $p > .20$].

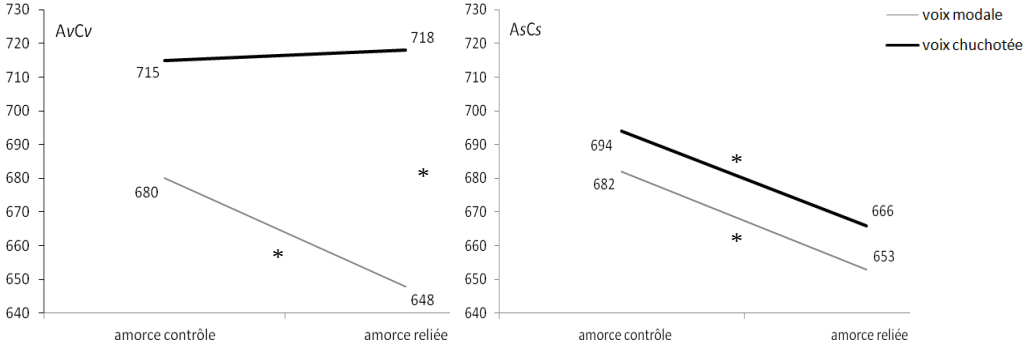


FIGURE 1 : Temps de réaction moyens (ms) par sujets en voix modale *vs* chuchotée selon l'amorce ; * différence significative par sujets et par items

4 Discussion

Chuchoter est un comportement vocal assez habituel (Cirillo, 2004), dont tout un chacun fait l'expérience quasi quotidiennement. Les modifications phonétiques des phonèmes issues de cette phonation sont très régulières, notamment sur le plan acoustique : ralentissement, perte d'énergie, balance spectrale plus aigüe, changement formantique, et bien sûr dévoisement (pour le français, Vercherand, 2010 ; Fux, 2012). Face à la voix modale, le chuchotement affecte de manière beaucoup plus critique les phonèmes voisés que les sourds et tout aussi systématiquement qu'une règle d'assimilation. Ainsi, même si elles ne relèvent pas d'une allophonie phonologique, les variantes chuchotées 'pragmatiques' (liées au contexte et à l'intension de communication) des phonèmes voisés sont tout autant déterminées.

Comme souligné dans l'Introduction, les obstruantes chuchotées, bien que modifiées phonétiquement, conserveraient des traces acoustiques de leur statut phonologique [±voisé]. Si l'auditeur se sert effectivement de ces traces pour reconnaître un mot chuchoté isolé, il peut le faire de façons très différentes selon la nature des représentations phonologiques lexicales. Dans une vision exemplariste où un même mot est représenté par une liste d'exemplaires construite par toutes les expériences auditives de ce mot, un auditeur pourrait donc aussi stocker des variantes chuchotées de mots. L'auditeur devrait ainsi pouvoir récupérer automatiquement les formes chuchotées des mots à consonne voisée (*désert*) tout aussi rapidement que celles à consonne sourde (*dessert*). Or, aucun effet d'amorçage sur la cible SABLE n'est observé si l'amorce est *désert*. L'auditeur n'accèderait donc pas immédiatement et automatiquement à un exemplaire stocké des mots chuchotés à consonne voisée.

Nos résultats sont par contre en accord avec une vision exemplariste moins stricte dans laquelle des prototypes de mots seraient stockés, le prototype d'un mot évoluant et s'alimentant au fur et à mesure de nos expériences avec celui-ci (Pierrehumbert, 2001). S'éloignant plus de sa forme modale qu'un mot chuchoté à consonne sourde, un mot chuchoté à consonne voisée constitue probablement un moins bon exemplaire du prototype, ce qui rendrait ainsi compte de l'absence d'effet d'amorçage dans le cas d'une amorce chuchotée à consonne voisée.

Nos résultats semblent aussi être en accord avec des modèles abstractionnistes de la reconnaissance des mots parlés, où les mots sont stockés sous la forme de séquences de segments discrets en termes de traits (Marslen-Wilson, Warren, 1994) ou de phonèmes (McClelland, Elman, 1986). Le signal de parole serait dans un premier temps converti en une séquence de segments discrets, écartant ainsi tous les détails acoustiques non pertinents pour l'identification. Le résultat de ce traitement serait ensuite projeté sur les représentations symboliques abstraites stockées en mémoire. Dans ces modèles, le caractère voisé/non voisé est alors extrait et/ou reconstruit à un niveau pré-lexical de traitement, avant le contact avec le lexique mental. La réalisation chuchotée des obstruantes sourdes étant assez proche de leur réalisation modale, la reconnaissance du trait [-vois] est probablement immédiate, ce qui expliquerait un accès rapide à la forme abstraite de *dessert*, facilitant le traitement subséquent du mot cible CHOCOLAT qui lui est sémantiquement relié. L'absence d'effet d'amorçage d'un mot chuchoté à obstruante voisée (*désert*) va quant à lui dans le sens d'un défaut de reconnaissance immédiate ou d'une reconstruction plus difficile du trait [+voisé] en parole chuchotée. Les mots chuchotés à obstruante voisée resteraient ambigus au moment où le mot cible est présenté, et aucun effet d'amorçage n'est alors observé.

Des analyses additionnelles semblent également montrer que la présentation d'amorces chuchotées, en comparaison à des amorces non chuchotées, a eu comme conséquence globale d'augmenter les temps de réaction sur les mots cibles. Cette augmentation des temps de réaction, lorsque les amorces sont chuchotées en comparaison à des amorces non chuchotées, s'est révélée n'être pleinement significative que pour les amorces reliées comportant une obstruante voisée. Un tel résultat semble également soutenir l'argument qu'au moment de l'écoute des mots cibles l'ambiguïté liée à la spécification [+voisé] de l'obstruante chuchotée n'est pas résolue, ce qui a pour conséquence d'augmenter plus spécifiquement les temps de réaction (70 ms) sur les mots cibles reliés à un mot amorce à obstruante voisée (i.e. *désert*–SABLE).

5 Conclusion

En ligne avec les observations de Fux (2012), nos résultats infirment une reconnaissance automatique et immédiate des obstruantes voisées chuchotées en français, comme cela a pu être suggéré par Vercherand (2010) et par Meynadier et al. (2013). Ainsi, même si des études montrent que les obstruantes voisées totalement assourdies, soit par chuchotement soit par assimilation, maintiendraient des traces phonétiques de leur identité sous-jacente, en phonation chuchotée ces consonnes resteraient malgré tout ambiguës au moins un temps pour l'auditeur.

Dans la mesure où, dans notre étude, les mots cibles étaient présentés dès la fin des amorces chuchotées, une hypothèse alternative est que la reconstruction du trait de voisement en parole chuchotée prend un certain temps. Des analyses complémentaires du délai entre la fin de présentation des amorces et le début d'affichage des cibles sont nécessaires de façon à obtenir de plus amples informations quant au décours temporel des processus impliqués dans la reconnaissance du trait de voisement en parole chuchotée.

Références

ANDRÉ C., GHIO A., CAVÉ C., TESTON B. (2003). PERCEVAL: a Computer-Driven System for Experimentation on Auditory and Visual Perception. *Proceedings of the 15th International Conference on Phonetic Sciences*, 1421-1424. Barcelona.

CIRILLO J. (2004). Communication by unvoiced speech: the role of whispering. *Anais Da Academia Brasileira de Ciências* 76(2), 413-423.

CREVIER-BUCHMAN L., VAISSIÈRE J., HENRICH N., VINCENT C., HANS S., BRASNU D. (2009). Laryngeal behavior in whispered voice: a study using high speed imaging. *The Voice Foundation's 38th Annual Symposium: Care of the Professional Voice*. Philadelphia.

FERRAND L., ALARIO F.-X. (1998). Normes d'association verbales pour 366 noms d'objets concrets. *L'Année Psychologique* 98, 659-709.

FUX T. (2012). *Vers un système indiquant la distance d'un locuteur par transformation de sa voix*. Thèse de doctorat. Université de 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.

GHIO A., TESTON B. (2004). Evaluation of the acoustic and aerodynamic constraints of a pneumotachograph for speech and voice studies. *Proceedings of the. 4th International Conference on Voice Physiology and Biomechanics*, 55-58. Marseille

HALLÉ P., ANDROJNA K., SEGUÍ J. (2012). L'assimilation de voisement en français : elle vaut pour les non-mots autant que les mots. *Actes des 29è Journées d'Etude sur la Parole*, 441–448. Grenoble.

MEYNADIER Y. (2015). Aerodynamic tool for phonology of voicing. *USB Proceedings of the 18th International Conference on Phonetic Sciences*, paper#0497. Glasgow.

MEYNADIER Y., DUFOUR S., GAYDINA, Y. (2013). Duration as perceptual voicing cue in whisper. *Proceedings of the 6th Phonetics and Phonology in Iberia Conference*, poster. Lisbon.

MEYNADIER Y., GAYDINA, Y. (2013). Aerodynamic and durational cues of phonological voicing in whisper. *Proceedings of the 14th Interspeech*, 335-339. Lyon.

MALÉCOT A., PEEBLES K. (1965). An optical device for recording glottal adduction-abduction during normal speech. *Zeitschrift Für Phonetik, Sprachw. Und Kommunikationsf.* 18, 545-550.

MARSLEN-WILSON W. D., WARREN P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review* 101, 653-675.

MCCLELLAND J. L., ELMAN J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology* 18, 1– 86.

PIERREHUMBERT J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee and P. Hopper (eds.) *Frequency effects and the emergence of lexical structure*. John Benjamins, Amsterdam, 137-157.

SNOEREN N., SEGUÍ J., HALLÉ P. (2008). On the role of regular phonological variation in lexical access: Evidence from voice assimilation in French. *Cognition* 108(2), 512–521.

TABOSSI P. (1996). Cross-Modal Semantic Priming. *Lang. and Cognitive Processes* 11(6), 569-576.

VERCHERAND G. (2010). *Production et perception de la parole chuchotée en français : analyse segmentale et prosodique*. Thèse de doctorat. Université Paris VII.

Acquisition et reconnaissance automatique d'expressions et d'appels vocaux dans un habitat.

Michel Vacher[1] Benjamin Lecouteux[2] Frédéric Aman[1]
François Portet[2] Solange Rossato[2]
(1) CNRS, LIG, F-38000 Grenoble, France
(2) Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
Michel.Vacher@imag.fr, Benjamin Lecouteux@imag.fr, Frederic.Aman@imag.fr,
francois.Portet@imag.fr, Solange.Rossato@imag.fr

RÉSUMÉ
Cet article présente un système capable de reconnaître les appels à l'aide de personnes âgées vivant à domicile afin de leur fournir une assistance. Le système utilise une technologie de Reconnaissance Automatique de la Parole (RAP) qui doit fonctionner en conditions de parole distante et avec de la parole expressive. Pour garantir l'intimité, le système s'exécute localement et ne reconnaît que des phrases prédéfinies. Le système a été évalué par 17 participants jouant des scénarios incluant des chutes dans un Living lab reproduisant un salon. Le taux d'erreur de détection obtenu, 29%, est encourageant et souligne les défis à surmonter pour cette tâche.

ABSTRACT
Acquisition and recognition of expressions and vocal calls in a smart home

This paper presents a system to recognize calls for help in the home of seniors to provide reassurance and assistance. The system is using an ASR which must operate with distant and expressive speech. Moreover, privacy is ensured by running the decoding on-site and not on a remote server. Furthermore the system was biased to recognize only set of sentences. The system has been evaluated in a smart space reproducing a typical living room where 17 participants played scenarios including falls. The results showed a promising error rate, 29%, while emphasizing the challenges of the task.

MOTS-CLÉS : traitement automatique de la parole, parole expressive, habitat intelligent.

KEYWORDS: automatic speech analysis, expressive speech, smart home.

1 Introduction

Le vieillissement rapide de la population des pays industrialisés représente l'un des défis majeurs du 21e siècle. En effet, en France, on estime que le nombre de Personnes Âgées (PA) de plus de 60 ans représentera 28,4% de la population en 2020 (9,4% auront plus de 75 ans) et 32,6% en 2060 (16,2% auront alors plus de 75 ans) (Blanpain & Chardon, 2010). Ceci est à mettre en relation avec l'espérance de vie sans incapacité qui a diminué pour être 61,9 ans en France en 2010, ce qui reste dans la moyenne de l'UE (INED, 2012). Lorsqu'une PA perd son autonomie, l'assistance d'un tiers devient nécessaire, celui-ci est généralement désigné sous le terme d'« aidant ». Ce rôle est en fait souvent tenu par un ensemble d'acteurs représentés majoritairement par la famille proche, en général le conjoint ou les enfants, qui doivent assumer les soins. Par ailleurs, cette augmentation du nombre de retraités aura un impact très important sur la société et les finances publiques à travers les régimes de retraite et la sécurité sociale (EPC, 2003). Le défi posé dès maintenant à notre société est de permettre à nos aînés de pouvoir vivre de façon autonome et confortable aussi longtemps que possible en toute quiétude alors même que le nombre de jeunes pouvant contribuer à leur support sera en constante

diminution.

Avec l'âge, le risque de chute croît. Environ 30% des personnes de plus de 65 chutent chaque année et ces chutes peuvent engendrer un fort taux de mortalité, morbidité et de souffrance pour les personnes âgées et leurs familles, elles augmentent par ailleurs la pression sur le coût social induit par une hospitalisation ou les soins médicaux (WHO Regional Office for Europe, 2004). D'autres phénomènes tels que les risques de crise cardiaques, arthrose etc. peuvent être des facteurs réduisant l'autonomie des personnes âgées et conduisent à un placement en institut spécialisé alors même que les personnes concernées peuvent encore posséder un degré d'autonomie suffisant pour vivre chez elles. L'une des principales raisons de ce placement restant la crainte d'un accident non signalé lorsque la personne est seule chez elle.

La détection des situations de détresse telles que les chutes, les immobilisations involontaires ou encore les évanouissements est un enjeu important pour soutenir la vie autonome des personnes âgées. Une solution communément proposée s'articule autour de capteurs cinématiques portés par la personne, ce qui constitue une contrainte dans la vie de tous les jours, avec des risques d'oubli voire même le refus de les porter car il sont jugés stigmatisants ou sans bénéfice immédiat (Bloch *et al.*, 2011). L'interface vocale peut constituer une alternative car ce type de technologie a atteint la maturité suffisante pour être utilisé ici tout en libérant la contrainte du port permanent des capteurs sur soi (Portet *et al.*, 2013). Un exemple type est celui d'un système de dialogue adapté à la personne. Par exemple, (Hamill *et al.*, 2009) ont développé un système PERS (*Personal Emergency Response System*) de dialogue vocal pour réagir à un appel spécifique d'urgence et décider de la réponse à apporter par une série de réponses fermées (« oui » et « non »). Un autre exemple est apporté par le projet ROBOCARE (Bahadori *et al.*, 2004) dans lequel un robot roulant comportant un écran d'ordinateur affichant un visage animé (avatar) a été conçu pour interagir spontanément avec l'utilisateur afin de lui signaler un danger ou répondre à une question.

Dans cet article nous présentons une approche permettant d'assister des personnes âgées dans leur maison par l'identification automatique d'appels vocaux, notamment dans le cas de situations de détresse où l'usage de la parole est encore possible. Une partie des résultats sont présentés dans (Vacher *et al.*, 2015c) ; la seconde partie résulte d'expériences effectuées sur un nouveau corpus. Les défis à relever pour rendre une telle application possible sont nombreux (Vacher *et al.*, 2011). Premièrement le cas d'application consiste à reconnaître de la parole distante puisque les microphones ne sont pas portés par la personne. Ceci implique de la parole atténuée, potentiellement réverbérée voire bruitée par d'autres sources sonores. Deuxièmement, la voix d'intérêt dans cette application présente de nombreuses différences avec la voix typique qui est celle des applications de Reconnaissance Automatique de la Parole (RAP) grand public, ces différences sont introduites par les modifications dues aux effets de l'âge du locuteur, à la situation fortement émotive dans laquelle il se trouve ainsi que l'occurrence de certaines pathologies. À ceci s'ajoute le manque de connaissance et de corpus réels sur l'appel de détresse. Les quelques études sur le sujet ont porté sur les urgences téléphoniques (Lefter *et al.*, 2011; Demenko & Jastrzebska, 2012) qui n'incluent pas nécessairement de personnes âgées.

Dans notre étude nous nous intéressons au cas particulier de la parole distante et expressive. L'objectif de l'étude est de mettre en place un Système de Reconnaissance Automatique de la Parole (SRAP) à l'état de l'art et de tester ses performances sur ce type de parole. Nous nous plaçons dans la situation où une PA est seule chez elle et possède un système de télé-lien social utilisant un microphone sans qu'aucun capteur ne soit porté par cette personne. Afin de recueillir un corpus d'étude réaliste, nous avons enregistré un ensemble de participants représentatifs de notre cible (au niveau de l'âge) simulant des situations de détresses dans le Living lab du laboratoire. Ces situations ont été observées lors d'une étude socio-ethnographique. En effet, enregistrer de la véritable voix de détresse dans un habitat utilisable pour une expérience reste un défi méthodologique et éthique. Dans un souci de respect de la vie privée, nous privilégions une approche *in situ* afin que la parole ne soit pas traitée par un serveur distant. Par ailleurs, l'utilisation d'un vocabulaire restreint assure la certitude que l'identification se limite aux appels vocaux. Cet article est organisé comme suit : la section 2 introduit les méthodes d'acquisition et de reconnaissance de la parole au sein de l'habitat ainsi que la méthode de détection

des appels à l'aide. La section 3 présente les expériences et finalement les résultats sont discutés dans la section 4.

2 Méthode

Pour mettre en place un système de reconnaissance d'appels vocaux en situation de détresse, la première étape a consisté à étudier ces situation sur le terrain et à définir leur environnement typique ainsi que le lexique utilisé. Cette étude est utile pour spécifier le système mais aussi pour mettre en place des scénarios de collecte de corpus. Le système global de reconnaissance d'appels vocaux est présenté avant de développer la modélisation acoustique et la méthode de reconnaissance.

2.1 Situations de détresse domestiques

La reconnaissance d'appels à l'aide s'effectue dans le contexte d'une maison équipée du dispositif *e-lio* [1], système de télé-lien social permettant de déclencher des appels (audio, vidéo ou urgence) entre les personnes âgées et leurs proches à l'aide d'une télécommande. Un objectif du projet était d'ajouter une commande vocale à ce système. Les paramètres de l'expérience ainsi que les situations de détresse ont été élaborés à partir d'une étude sociologique du laboratoire GRePS (Bobillier Chaumon *et al.*, 2013) auprès d'un grand nombre de personnes âgées qui a montré que cet équipement doit être installé sur une table de la pièce de vie, face au canapé et à la télévision. Ainsi une alerte peut facilement être lancée si la personne tombe à cause du tapis ou n'arrive pas à se lever du canapé. Plus de détails sont donnés dans l'article (Bouakaz *et al.*, 2014).

Ces études ont permis de spécifier le contexte des chutes, les mouvements pendant les chutes ainsi que la réaction des personnes une fois au sol. Les phrases prononcées pendant et après la chute ont également été identifiées : "*Ah, zut, qu'est-ce qui m'arrive ? Oh merde, merde !*".

2.2 Système d'analyse sonore en ligne

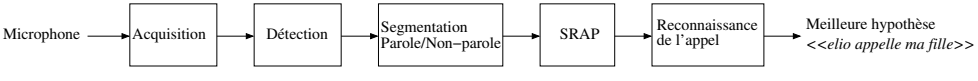


FIGURE 1: Architecture du système d'analyse sonore CIRDOX

Le traitement de l'audio est réalisé par le logiciel CIRDOX(Aman *et al.*, 2016). La Figure 1 décrit le pipeline de traitement. Le flux audio est acquis en continu et les événements sonores sont détectés à la volée en utilisant une décomposition en ondelettes associée à un seuillage adaptatif (Vacher *et al.*, 2004). Les événements sonores sont alors classifiés comme non-parole ou parole et dans ce cas envoyés au SRAP qui les transmet au système de reconnaissance des appels à l'aide.

Dans cet article, nous nous focalisons sur le SRAP et présentons différentes stratégies visant à améliorer la reconnaissance des appels à l'aide, la modélisation acoustique ainsi que l'étape de reconnaissance.

1. http ://www.technosens.fr/

2.3 Modélisation langagière

La modélisation acoustique est basée sur des Modèles de Markov Cachés (MMC) gauche-droite à trois états dépendants du contexte. Les paramètres acoustiques sont basés sur des MFCC à 40 paramètres (13 coefficients + delta + delta delta + énergie). Au final les modèles acoustiques comportent 11000 états partagés dépendants du contextes avec un total de 150000 gaussiennes. Par ailleurs une adaptation SAT+fMLLR est appliquée (Povey *et al.*, 2011b).

Les modèles ont été estimés sur 500 heures de français annotées, issues d'émissions radiophonique ou TV (ESTER 1 + 2 + REPERE) (Gravier *et al.*, 2004) et 7 heures précédemment enregistrées dans un appartement intelligent (Vacher *et al.*, 2014) qui regroupe 60 locuteurs interagissant avec l'appartement et 28 minutes du corpus "voix-détresse" (Aman, 2014) composé d'enregistrements de locuteurs énonçant des appels de détresse.

2.3.1 Modèles acoustiques « Subspace GMM »

Les modèles à base de mélanges de gaussiennes et les modèles à sous-espaces sont utilisés pour le calcul d'émission des probabilités au sein des états du MMC, avec la particularité pour les SGMM d'avoir des moyennes et poids générés à partir de sous-espaces de mélanges de gaussiennes, via une projection pondérée.

Les modèles SGMM introduits par (Povey *et al.*, 2011a) sont décrits par l'équation suivante :

$$p(\mathbf{x}|j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \mu_{jmi}, \mathbf{\Sigma}_i) \text{ avec } \mu_{jmi} = \mathbf{M}_i \mathbf{v}_{jm} \text{ et } w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \mathbf{v}_{jm}}.$$

où $\mathbf{x}$ correspond au vecteur de paramètres, $j \in \{1..J\}$ est un état du MMC, i est l'index de la gaussienne, m est un sous-état de j et c_{jm} le poids du sous-état m. À chaque état j sont associés un vecteur $\mathbf{v}_{jm} \in \mathbb{R}^S$ (S est la dimension du sous-espace phonétique), les moyennes μ_{jmi}, la pondération des mélanges w_{jmi} et un nombre de Gaussiennes partagé I. Le sous-espace phonétique $\mathbf{M}_i$, le poids de projection $\mathbf{w}_i^T$ et la matrice de covariance $\mathbf{\Sigma}_i$, c.-à-d. les paramètres globaux $\mathbf{\Phi}_i = \{\mathbf{M}_i, \mathbf{w}_i^T, \mathbf{\Sigma}_i\}$, sont communs à tous les états. Ces paramètres peuvent être partagés et estimés à travers différentes conditions d'enregistrement.

Un mélange générique de gaussiennes I appelé *Universal Background Model (UBM)* est estimé sur toutes les données d'apprentissage, et sert à l'initialisation du SGMM.

Dans nos expériences nous avons estimé séparément trois UBM contenant 1K gaussiennes à partir des données SWEET-HOME (7h), Voix-détresse (28mn) et ESTER+REPERE (500h). Ces trois UBM ont été fusionnés en un seul (en fusionnant les paires les plus proches (Zouari & Chollet, 2006)) ; le modèle final est alors spécifiquement biaisé par les enregistrements dans la maison ainsi que par la parole expressive.

2.3.2 Modèles de langage

Cet article est focalisé sur la reconnaissance d'appels à l'aide. Pour ce faire, les phrases du corpus AD80 (Aman *et al.*, 2013) ont servi de base pour développer le modèle de langage spécialisé du système. Ce corpus a été enregistré par 43 personnes âgées et 52 non âgées dans notre laboratoire ainsi que dans une maison de retraite, avec pour objectif d'étudier la reconnaissance de la parole pour les personnes âgées. Ce corpus est composé pour chaque locuteur de 81 phrases classiques, 31 commandes vocales pour la domotique et 58 phrases d'appel à l'aide.

Quelques unes de ces phrases sont données dans la table 1. Les phrases d'appel identifiées dans l'étude reportée à la section 2.1 ont été incluses dans AD80.

2.4 Reconnaissance des appels à l'aide

Comme dit précédemment, la voix en situation de détresse peut être assez atypique et perturber le SRAP. Cependant, dans de nombreuses situations, un groupe de mots mal décodé est phonétiquement proche d'un segment "bon candidat". C'est pourquoi notre approche consiste à transcrire chaque appels possibles et l'hypothèse du SRAP en un graphe de phonèmes dans lequel chaque chemin correspond à une variante phonétique. Cette approche permet de prendre en considération un certain nombre d'erreurs de reconnaissance comme des fins de mots erronées ou des variations légères. Ainsi, chaque hypothèse du SRAP H_i est phonétisée et chaque commande vocale T_j est alignée à l'hypothèse H_i en utilisant une distance de Levenshtein. Le nombre de phrases prédéfinies est de 50 et correspondent à 11 actions différentes.

La décision de sélectionner ou non une séquence détectée est alors prise en fonction d'un seuil défini empiriquement dépendant du nombre de phonèmes alignés pour chaque candidat. À partir de là, le taux d'erreur domotique (TED) est défini par la formule : $\text{TED} = \frac{\text{Nombre d'appels manqués}+\text{Nombre de fausses alarmes}}{\text{Nombres d'appels}}$

3 Expériences et résultats

3.1 Expériences actées en situation réaliste : le corpus de parole CIRDO

Les enregistrements ont été réalisé dans une pièce du Living lab du LIG équipée comme décrit précédemment. Le protocole ainsi que le corpus CIRDO obtenu sont détaillés dans (Vacher *et al.*, 2016). Les scénarios comportaient des appels à l'aide lors de chute ou en cas de blocage de hanche. Tous les enregistrements audio ont été transcrits manuellement en utilisant *Transcriber* (Barras *et al.*, 2001) et les segments audio ont alors été extraits pour analyse. Durant le scénario, certains participants soupiraient, gémissaient, toussaient, criaient ou se raclaient la gorge : ces sons n'ont pas été transcrits. Dans le même ordre d'idée, les paroles mélangées avec les bruits de chute ont été ignorées. À la fin, chaque locuteur avait prononcé entre 10 et 65 phrases ou interjections (("*ah*", "*oh*", "*aïe*", etc.) (Table 2). Certaines phrases étaient proches de celles prévues dans les scénarios ("*je peux pas me relever*", "*e-lio appelle du secours*", etc.), tandis que d'autres différaient ("*oh bein on est bien là tiens* "). Dans la pratique, les participants font de longues pauses au milieu d'une phrase ou prononcent des phrases spontanées, utilisent des interjections ou des sons non-verbaux.

3.2 Reconnaissance « off-line »des appels sur le corpus CIRDO

Le SRAP utilisé est basé sur Kaldi (Povey *et al.*, 2011b). Le modèle SGMM présenté dans la section 2.3 a été utilisé comme modèle acoustique. Le modèle de langage générique a été estimé sur des corpus

Appels à l'aide	*Commandes domotiques*	*Phrases classiques*
Aïe aïe aïe ⋆	Appelle quelqu'un e-lio ⋆	Bonjour madame
Oh là ⋆	e-lio, appelle quelqu'un ⋆	Ça va très bien
Merde ⋆	e-lio tu peux appeler une ambulance	Où sont mes lunettes
Je suis tombé ⋆	e-lio tu peux téléphoner au SAMU	Le café est brûlant
Je peux pas me relever ⋆	e-lio, appelle du secours	J'ai ouvert la porte
Qu'est-ce qu'il m'arrive ⋆	e-lio appelle les secours	Je me suis endormi tout de suite
Aïe ! J'ai mal ⋆	e-lio appelle ma fille	Il fait soleil
Oh là ! Je saigne ! Je me suis blessé ⋆	e-lio appelle l'infirmière	Ce livre est intéressant
Aidez-moi	e-lio appelle le SAMU !	Je dois prendre mon médicament
Au secours	e-lio appelle les pompiers !	J'allume la lumière

TABLE 1: Exemples extraits du corpus AD80 (les ⋆ correspondent à des phrases observées dans l'étude ethnographique)

Loc.	Age	Sexe	Nb. d'interjections ou phrases courtes		Loc.	Age	Sexe	Nb. d'interjections ou phrases courtes	
			Total	*# Appel*				*Total*	*# Appel*
S01	30	M	22	14	S10	16	M	19	15
S02	-	-	-		S11	52	M	12	12
S03	24	F	16	15	S12	28	M	15	12
S04	83	F	65	53	S13	66	M	24	21
S05	29	M	24	21	S14	52	F	23	2
S06	64	F	23	19	S15	23	M	20	19
S07	61	M	23	21	S16	40	F	29	27
S08	44	M	25	15	S17	40	F	24	21
S09	16	M	32	21	S18	25	F	17	14
Total	40.76		413	341					

TABLE 2: Composition du corpus audio CIRDO

Loc.	TEM (%)		TED	Loc.	TEM (%)		TED
	Total	*# Appel*			*Total*	*# Appel*	
S01	45,0	39,1	27,8	S11	21,3	17,0	16,7
S03	41,4	44,4	40,0	S12	30,8	25,0	25,0
S04	51,9	49,6	34,0	S13	45,9	43,6	23,8
S05	19,1	15,4	14,3	S14	67,0	54,8	50,0
S06	39,2	34,3	26,3	S15	21,5	19,5	5,3
S07	21,2	20,3	28,6	S16	14,9	11,76	7,4
S08	61,8	50,8	20,0	S17	21,4	22,4	19,0
S09	49,4	41,2	33,3	S18	57,7	44,9	71,4
S10	24,5	22,4	14,3				
Total	39,3	34,0	26,8				

TABLE 3: Taux d'erreur mot et domotique pour chaque participant

de journaux ainsi que sur Gigaword. Le lexique est d'environ 13K mots. Pour réduire la variabilité linguistique, ce modèle générique a été interpolé avec un modèle de langage lié au domaine se basant sur les scénarios. Le modèle de langage final est le résultat d'une interpolation avec 10% pour le modèle générique et 90% pour le modèle spécialisé. Cette interpolation a montré des résultats corrects dans nos précédentes expérimentations (Lecouteux *et al.*, 2011). L'avantage de cette interpolation permet de biaiser le modèle pour reconnaître des situations d'appel à l'aide tout en réduisant la reconnaissance de faux positifs.

Les résultats sur les données manuellement annotées sont donnés dans la table 3. Si nous considérons uniquement les appels à l'aide, le Taux d'Erreur de Mot (TEM) est 34% tandis qu'il monte à 39.3% quand toutes les interjections et phrases sont prises en considération. Malheureusement, le corpus utilisé ne permet pas de déterminer le taux de fausses alarmes car il ne contient pas de scénario comportant des appels qui ne soient pas des appels à l'aide. Nos précédentes études basées sur le corpus AD80 montrent un rappel, précision et F-mesure d'environ 88,4 %, 86,9 % and 87,2 % (Aman *et al.*, 2013). Cependant, ce corpus a été enregistré dans des conditions très différentes de celles d'un studio, comme l'avait été CIRDO. En moyenne, le TED est 26,8 %.

4 Discussion

Si nous comparons ces résultats avec ceux obtenus en utilisant le corpus de parole lue AD80, ils sont nettement inférieurs, le TEM est 26,8% contre 14,5% (Aman *et al.*, 2013). Ceci peut s'expliquer par des différences notables entre les conditions d'enregistrement des 2 corpus :

— AD80 est basé sur des enregistrements de locuteurs qui sont confortablement assis et en condition proche face au microphone, tandis que CIRDO l'a été en condition distante,

— le corpus CIRDO a été enregistré avec des participants qui tombent sur le sol ou se trouvent bloqués sur le canapé. De plus, ils ont été encouragés à se mettre en situation et parler avec l'émotion qu'ils auraient ressentie dans ce type de situation. Les enregistrements contiennent

donc une parole expressive, mais ceci ne garantit pas totalement qu'elle l'aurait été de cette manière en condition réelle.

Par contre, si nous comparons maintenant ces résultats avec ceux obtenus en ligne avec CIRDOX et un SRAP utilisant Sphinx3 (Lee *et al.*, 1990) et des modèles acoustiques de type GMM adaptés au locuteur, nous observons une amélioration notable (Vacher *et al.*, 2015a). En effet le TEM était 49,5% (contre 34%) et le TED 33% (contre 26,8%), ceci semble indiquer que des modèles plus élaborés, et qui prennent en compte plus de données d'apprentissage comme les sGMM, permettent une amélioration sensible des résultats de reconnaissance sur de la parole enregistrée en conditions difficile.

Si l'on observe le taux d'erreur au niveau des appels (TED), on observe qu'il est de 26.8% ; de plus, à l'exception d'un seul locuteur, le TED est toujours en dessous de 50% et inférieur à 20% pour 6 locuteurs. Cela suggère qu'un appel à l'aide a de meilleures chances d'être détecté si le locuteur est en capacité de le répéter deux ou trois fois. Cependant, si le système n'identifie pas le premier appel de détresse à cause de la voix altérée par l'émotion, il y a des chances que cette non-détection augmente encore l'émotion ressentie par la personne, ce qui aura pour conséquence une difficulté accrue pour le système de reconnaissance. Dans le même ordre d'idée, ce corpus a été enregistré en conditions réalistes, mais ne reproduisant pas totalement la réalité : la production vocale des personnes âgées fragiles est difficilement reproductible par des personnes jeunes. Il convient d'éviter au maximum la nécessité pour les personnes d'avoir à répéter un appel, il est donc important de poursuivre les efforts en vue de résoudre ces erreurs.

5 Conclusion et perspectives

Cette étude s'est focalisée sur la RAP dans le cadre de maisons intelligentes et en conditions distantes et réalistes : des conditions très différentes de celles d'un corpus enregistré assis et proche du micro. En effet, le corpus CIRDO constitué d'appels de détresse en parole distante et qui simulent des conditions réalistes (chutes, blocage sur un canapé) a été utilisé. Le TEM obtenu en sortie d'un système SRAP dédié était 36,3% au niveau des appels de détresse. Cependant, grâce à un filtrage des hypothèse au niveau phonétique, plus de 70% des appels ont pu être détectés.

Ces résultats obtenus en conditions réalistes donnent une bonne idée des performances qui peuvent être obtenues avec un SRAP état de l'art dans ces conditions spécifiques avec des utilisateurs finaux. Ces résultats on été obtenus dans des cas de situations de détresse, perturbées par les émotions ; ce type d'expérimentation pourra être étendu à d'autres situations où l'expressivité est particulièrement marquée.

Comme expliqué précédemment, les résultats obtenus doivent être améliorés afin que le système puisse être utilisé en « production ». Deux axes peuvent être explorées, le premier pourrait être une adaptation des modèles acoustiques à la prosodie de la parole expressive. L'enregistrement du type de corpus nécessaire à cette adaptation est délicat car il doit se faire en conditions réelles, ce qui représente un inconvénient majeur. Une autre piste serait la reconnaissance de répétitions à intervalles réguliers d'évènements phonétiquement similaires : même si la reconnaissance de la parole en tant que telle n'est pas bonne, cela peut être signe d'un problème à régler rapidement.

Ces travaux montrent la nécessité de procéder à des analyses basées sur des corpus enregistrés en conditions écologiques. Il sont complémentaires à d'autres travaux sur la commande vocale de la domotique (Vacher *et al.*, 2015b) pour lesquels l'aspect parole expressive n'était pas prédominant mais qui ont montré la nécessité d'une adaptation au vocabulaire et aux tournures propres de chaque utilisateur au moyen d'une analyse lexicale.

Références

AMAN F. (2014). *Reconnaissance automatique de la parole de personnes âgées pour les services d'assistance à domicile*. PhD thesis, Université de Grenoble, Ecole doctorale MSTII.

AMAN F., VACHER M., PORTET F., DUCLOT W. & LECOUTEUX B. (2016). CirdoX : an On/Off-line Multisource Speech and Sound Analysis Software. In *LREC 2016*. (Accepted Paper).

AMAN F., VACHER M., ROSSATO S. & PORTET F. (2013). Speech Recognition of Aged Voices in the AAL Context : Detection of Distress Sentences. In *The 7th International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2013*, p. 177–184, Cluj-Napoca, Romania.

BAHADORI S., CESTA A., GRISETTI G., IOCCHI L., LEONE R., NARDI D., ODDI A., PECORA F. & RASCONI R. (2004). RoboCare : Pervasive Intelligence for the Domestic Care of the Elderly. *Intelligenza Artificiale*, **1**(1), 16–21.

BARRAS C., GEOFFROIS E., WU Z. & LIBERMAN M. (2001). Transcriber : development and use of a tool for assisting speech corpora production. *Speech Communication*, **33**(1-2), 5–22.

BLANPAIN N. & CHARDON O. (2010). Projections de population à l'horizon 2060 : Un tiers de la population âgé de plus de 60 ans. Institut national de la statistique et des etudes economiques (France). [in French].

BLOCH F., GAUTIER V., NOURY N., LUNDY J., POUJAUD J., CLAESSENS Y. & RIGAUD A. (2011). Evaluation under real-life conditions of a stand-alone fall detector for the elderly subjects. *Annals of Physical and Rehabilitation Medicine*, **54**, 391–398.

BOBILLIER CHAUMON M., CROS F., CUVILLIER B., HEM C. & CODREANU E. (2013). Concevoir une technologie pervasive pour le maintien à domicile des personnes âgées : la détection de chutes dans les activités quotidiennes. In *Activités Humaines, Technologies et bien-être, Congrès EPIQUE (Psychologie Ergonomique)*, p. 189–199, Belgique - Bruxelles.

BOUAKAZ S., VACHER M., BOBILLIER-CHAUMON M.-E., AMAN F., BEKKADJA S., PORTET F., GUILLOU E., ROSSATO S., DESSERÉE E., TRAINEAU P., VIMON J.-P. & CHEVALIER T. (2014). CIRDO : Smart companion for helping elderly to live at home for longer. *Innovation and Research in BioMedical engineering (IRBM)*, **35**(2), 101–108.

DEMENKO G. & JASTRZEBSKA M. (2012). Analysis of voice stress in call centers conversations. In *Speech Prosody*.

EPC (2003). The impact of ageing populations on public finances : overview of analysis carried out at EU level and proposals for a future work programme. Economic Policy Committee, Union Européenne.

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* : European Language Resources Association.

HAMILL M., YOUNG V., BOGER J. & MIHAILIDIS A. (2009). Development of an automated speech recognition interface for personal emergency response systems. *Journal of NeuroEngineering and Rehabilitation*, **6**.

INED (2012). Dernières données sur l'espérance de vie sans incapacité des 27 pays de l'ue. Communiqué de Presse de l'Ined.

LECOUTEUX B., VACHER M. & PORTET F. (2011). Distant Speech Recognition in a Smart Home : Comparison of Several Multisource ASRs in Realistic Conditions. In *Proc. InterSpeech*, p. 2273–2276.

LEE K.-F., HON H.-W. & REDDY R. (1990). An overview of the SPHINX speech recognition system. *IEEE TASSP*, **38**(1), 35–45.

LEFTER I., ROTHKRANTZ L. J., VAN LEEUWEN D. A. & WIGGERS P. (2011). Automatic stress detection in emergency (telephone) calls. *International Journal of Intelligent Defence Support Systems*, **4**(2), 148–168.

PORTET F., VACHER M., GOLANSKI C., ROUX C. & MEILLON B. (2013). Design and evaluation of a smart home voice interface for the elderly — Acceptability and objection aspects. *Personal and Ubiquitous Computing*, **17**(1), 127–144.

POVEY D., BURGET L., AGARWAL M., AKYAZI P., KAI F., GHOSHAL A., GLEMBEK O., GOEL N., KARAFIÁT M., RASTROW A., ROSE R. C., SCHWARZ P. & THOMAS S. (2011a). The subspace gaussian mixture model—a structured model for speech recognition. *Computer Speech & Language*, **25**(2), 404 – 439.

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. (2011b). The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* : IEEE Signal Processing Society. IEEE Catalog No. : CFP11SRW-USB.

VACHER M., AMAN F., ROSSATO S. & PORTET F. (2015a). Development of Automatic Speech Recognition Techniques for Elderly Home Support : Applications and Challenges. In *HCII, ITAP 2015*, volume Part II of *LNCS 9194*, p. 341–353, Los Angeles, CA, United States.

VACHER M., BOUAKAZ S., BOBILLIER CHAUMON M.-E., AMAN F., KHAN R. A., BEKKADJA S., PORTET F., GUILLOU E., ROSSATO S. & LECOUTEUX B. (2016). The CIRDO corpus : comprehensive audio/video database of domestic falls of elderly people. In *LREC 2016*. (Accepted Paper).

VACHER M., CAFFIAU S., PORTET F., MEILLON B., ROUX C., ELIAS E., LECOUTEUX B. & CHAHUARA P. (2015b). Evaluation of a context-aware voice interface for Ambient Assisted Living : qualitative user study vs. quantitative system evaluation. *ACM Transactions on Accessible Computing*, **7**(issue 2), 5 :1–5 :36.

VACHER M., ISTRATE D. & SERIGNAT J. (2004). Sound detection and classification through transient models using wavelet coefficient trees. In S. LTD, Ed., *Proc. 12th European Signal Processing Conference*, p. 1171–1174, Vienna, Austria.

VACHER M., LECOUTEUX B., AMAN F., ROSSATO S. & PORTET F. (2015c). Recognition of Distress Calls in Distant Speech Setting : a Preliminary Experiment in a Smart Home. In *6th Workshop on Speech and Language Processing for Assistive Technologies*, p. 1–7, Dresden, Germany : SIG-SLPAT.

VACHER M., LECOUTEUX B., CHAHUARA P., PORTET F., MEILLON B. & BONNEFOND N. (2014). The Sweet-Home speech and multimodal corpus for home automation interaction. In *The 9th edition of the Language Resources and Evaluation Conference (LREC)*, p. 4499–4506, Reykjavik, Iceland.

VACHER M., PORTET F., FLEURY A. & NOURY N. (2011). Development of Audio Sensing Technology for Ambient Assisted Living : Applications and Challenges. *International Journal of E-Health and Medical Communications*, **2**(1), 35–54.

WHO REGIONAL OFFICE FOR EUROPE (2004). What are the main risk factors for falls amongst older people and what are the most effective interventions to prevent these falls ? World Health Organisation.

ZOUARI L. & CHOLLET G. (2006). Efficient gaussian mixture for speech recognition. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 4, p. 294–297.

Adaptation de la prononciation pour la synthèse de la parole spontanée en utilisant des informations linguistiques

Raheel Qader[1] Gwénolé Lecorvé[1] Damien Lolive[1] Pascale Sébillot[2]
(1) IRISA/Université de Rennes 1, 6, rue de Kerampont, 22300 Lannion, France
(2) IRISA/INSA Rennes, 263 Avenue Général Leclerc, 35000 Rennes, France
{raheel.qader,gwenole.lecorve,damien.lolive,pascale.sebillot}@irisa.fr

RÉSUMÉ

Cet article présente une nouvelle méthode d'adaptation de la prononciation dont le but est de reproduire le style spontané. Il s'agit d'une tâche-clé en synthèse de la parole car elle permet d'apporter de l'expressivité aux signaux produits, ouvrant ainsi la voie à de nouvelles applications. La force de la méthode proposée est de ne s'appuyer que sur des informations linguistiques et de considérer un cadre probabiliste pour ce faire, précisément les champs aléatoires conditionnels. Dans cet article, nous étudions tout d'abord la pertinence d'un ensemble d'informations pour l'adaptation, puis nous combinons les informations les plus pertinentes lors d'expériences finales. Les évaluations de la méthode sur un corpus de parole conversationnelle en anglais montrent que les prononciations adaptées reflètent significativement mieux un style spontané que les prononciations canoniques.

ABSTRACT

Pronunciation adaptation for spontaneous speech synthesis using linguistic information.

This paper presents a new pronunciation adaptation method which adapts canonical pronunciations to a spontaneous style. This is a key task in text-to-speech as those pronunciation variants bring expressiveness to synthetic speech, thus enabling new potential applications. The strength of the method is to solely rely on linguistic features and to consider a probabilistic machine learning framework, namely conditional random fields, to produce the adapted pronunciations. Features are selected in a first series of experiments, then combined in the backend experiments. Results on the Buckeye conversational English speech corpus show that adapted pronunciations significantly better reflect spontaneous speech than canonical ones.

MOTS-CLÉS : Adaptation de la prononciation, parole spontanée, synthèse de la parole.

KEYWORDS: Pronunciation adaptation, spontaneous speech, speech synthesis.

1 Introduction

Les variantes de prononciation de mots ou énoncés ne sont pas prises en compte par les lexiques et modèles de prononciation utilisés dans les systèmes actuels de synthèse de la parole. Cela limite alors leur capacité à produire des signaux expressifs, notamment pour retranscrire un style spontané. Pour résoudre ce problème, nous proposons une nouvelle méthode d'adaptation de la prononciation dont le but est d'imiter le style spontané de locuteurs individuels et de pouvoir, à terme, être intégrée dans un moteur de synthèse de la parole. La langue étudiée est l'anglais.

Sous l'angle de l'apprentissage automatique, la méthode proposée consiste à prédire pour chaque phonème d'une prononciation canonique si celui-ci doit être supprimé, remplacé, gardé tel quel ou

Phonèmes canoniques	/k ɑ n s ʌ **n** t ɹ eɪ **t ʌ** d · ɪ **n** · oʊ h **aɪ oʊ** /
Phonèmes réalisés	/k ɑ n s **n̩** _ t ɹ eɪ _ **ɪ** d · ɪ **ɾ̃** · oʊ h **ɑ ʌ** /

TABLE 1 – Prononciations réalisée et canonique pour la séquence de mots « *concentrated in Ohio* ».

complété par des phonèmes à insérer. Pour cela, cette méthode repose sur des Champs Aléatoires Conditionnels (CAC), dont l'utilisation est très répandue pour la phonétisation de mots ou d'énoncés (Wang & King, 2011; Illina *et al.*, 2011; Lecorvé & Lolive, 2015). Les CAC sont très utiles car ils permettent d'intégrer et combiner simplement un très large panel d'informations. Précisément, la force de notre méthode est de ne s'appuyer, en complément des phonèmes canoniques, que sur des informations linguistiques car aucune information acoustique n'est disponible au moment de la phonétisation d'un énoncé en synthèse de la parole.

Les travaux connexes en production de variantes de prononciation peuvent être résumés d'après le type de l'approche retenue et la nature des informations utilisées. Tout d'abord, diverses approches par apprentissage automatique ont déjà été utilisées : des arbres de décision (Fosler-Lussier *et al.*, 1999; Vazirnezhad *et al.*, 2009), des forêts aléatoires (Dilts, 2013), des réseaux de neurones (Chen & Hasegawa-Johnson, 2004; Karanasou *et al.*, 2013), des modèles de Markov cachés (Prahallad *et al.*, 2006) et des CAC (Karanasou *et al.*, 2013). Pour aller plus loin, d'autres travaux ont également proposé de combiner différentes techniques (Vazirnezhad *et al.*, 2009; Kolluru *et al.*, 2014). Il est malheureusement difficile de comparer ces travaux car ceux-ci partagent rarement les mêmes données ou la même tâche exacte. Quant aux informations utilisées, des caractéristiques acoustiques peuvent être extraites à partir de signaux de parole d'un style visé et prises en compte pour l'adaptation de prononciations (fréquence fondamentale, énergie, durée, débit de parole…) (Bates & Ostendorf, 2002; Bell *et al.*, 2009, 2003), tandis que des informations linguistiques peuvent être dérivées de textes (distinction entre mots-outils et mots pleins, probabilité des mots, informations syllabiques, accentuation lexicale dans certaines langues…) (Vazirnezhad *et al.*, 2009; Bell *et al.*, 2009, 2003). Récemment, Dilts (2013) a présenté une étude poussée sur la combinaison de ces deux types d'informations. Ce travail est proche du nôtre mais diffère dans le sens où la technique d'apprentissage automatique est différente et l'objectif était uniquement de réduire les prononciations. En complément, notons également que (Chen & Hasegawa-Johnson, 2004) a montré que les informations d'un phonème canonique doivent être enrichies par celles de leur voisinage pour aboutir à de meilleures adaptations. Enfin, il est important de noter que la plupart des travaux du domaine visent la reconnaissance automatique de la parole alors que les approches pour la synthèse sont encore rares et qu'aucune ne fait un usage aussi intensif que le nôtre des informations linguistiques.

Dans cet article, la section 2 décrit le corpus de parole Buckeye utilisé dans nos travaux ; la section 3 présente la méthode et le protocole expérimental ; la section 4 étudie en préambule des caractéristiques isolées avant que celles-ci ne soient combinées dans les expériences finales de la section 5.

2 Le corpus de parole Buckeye

Nous avons utilisé le corpus de parole conversationnelle Buckeye (Pitt *et al.*, 2005). Ce corpus, en anglais, consiste en 40 entretiens non préparés avec des locuteurs de l'Ohio, aux États-Unis, chaque entretien durant 1 heure. 20 entretiens ont été sélectionnés au hasard, les autres ayant été laissés de côté pour d'éventuels futurs travaux. Les signaux de parole sont fournis avec des transcriptions vérifiées manuellement : une transcription orthographique et deux transcriptions phonétiques, l'une correspondant aux phonèmes canoniques qui auraient dû être prononcés si le style avait été neutre,

Phonèmes

phonème canonique (40) • **position du phonème dans la syllable (20)** • **position inversée du phonème dans la syllabe (22)** • **phonème en début, milieu ou fin de mot (40)**

Syllabes

accentuation lexicale de la syllabe (24) • **partie de la syllabe (24)** • type de syllabe (18) • **position de la syllabe dans le mot (20)**

Mots

mot (40) • graphème (16) • **est-ce un mot vide (d'après une liste) ? (24)** • **fréquence du mot en anglais (22)** • fréquence du mot dans l'entretien (18) • fréquence de la racine en anglais (16) • fréquence de la racine dans l'entretien (19) • position du mot dans l'énoncé (2) • position inverse du mot dans l'énoncé (0) • numéro d'occurrence du mot dans l'entretien (0) • classe grammaticale (17) • longueur en graphèmes (16) • longueur en syllabes (17)

Énoncés

position de l'énoncé dans l'entretien (3) • position inverse de l'énoncé dans l'entretien (4)

TABLE 2 – Liste des caractéristiques (hormis les phonèmes réalisés). En gras, celles qui ont été conservées à l'issue de la phase de sélection. Entre parenthèses, le nombre de votes reçus (cf. section 4).

l'autre aux phonèmes effectivement réalisés par le locuteur dans un cadre de parole spontanée. Chaque locuteur représente environ 7 400 mots et 22 800 phonèmes. Les phonèmes canoniques et réalisés ont été alignés automatiquement. Il en découle que 30 % des phonèmes et 57 % des mots sont prononcés différemment de ce qui était attendu. L'exemple de la table 1 permet de constater à quel point les prononciations réalisées diffèrent généralement des prononciations canoniques.

Ces annotations ont été complétées par de nombreuses autres au moyen d'outils automatiques, conduisant à un total de 23 caractéristiques portant sur les phonèmes, syllabes, mots et énoncés. Leur détail est donné par la table 2. Afin d'être compatible avec l'emploi de CAC, les fréquences ont été catégorisées à masses de probabilité équivalentes en « fréquent », « moyen » et « rare ». Nous présentons maintenant la méthode en elle-même.

3 Présentation de la méthode et du protocole expérimental

Nous formalisons l'adaptation de la prononciation comme la prédiction d'une séquence de phonèmes réalisés à partir d'une séquence de phonèmes canoniques. Expérimentalement, la qualité d'une adaptation se mesure alors à son taux d'erreurs entre les phonèmes prédits, dits *adaptés*, et phonèmes effectivement réalisés dans le corpus.

Dans notre méthode, nous avons choisi d'utiliser des CAC, type de modèles particulièrement adéquat pour l'apprentissage sur des données séquentielles et symboliques. Pour construire au mieux ces modèles, nous avons cherché à ajouter de nouvelles informations en entrée en plus des seuls phonèmes canoniques. Les différentes pistes qui ont été explorées pour cela sont illustrées par la figure 1. Nous les présentons ci-dessous et introduisons les questions qui s'y rattachent.

1. Principalement, chaque phonème p_i à prédire dépend de n caractéristiques $\{c_i^1, \ldots, c_i^n\}$, par exemple le phonème canonique à adapter, sa position dans la syllabe ou la fréquence du mot qui le contient. La question est alors de savoir quelles caractéristiques parmi toutes celles considérées sont pertinentes et quelles autres dégradent l'adaptation.

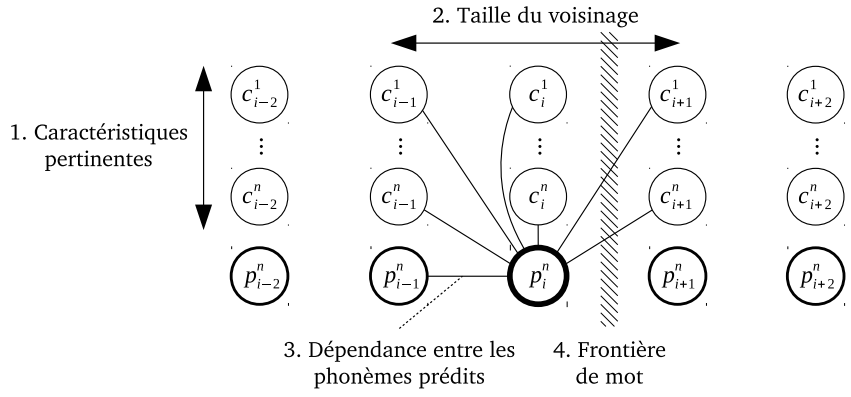


FIGURE 1 – Vue d'ensemble des dépendances et paramètres à traiter pour l'apprentissage des CAC. Les nœuds et arrêtes représentent respectivement différentes informations et leurs liens de dépendance.

2. Ensuite, le panel des informations peut être étendu au voisinage de p_i, par exemple en considérant également le phonème canonique précédent et le suivant, ainsi que les caractéristiques qui leur sont associées. Cette notion de voisinage se définit en pratique par une fenêtre de taille W autour de p_i. Le choix de cette taille est un point que nous avons étudié.
3. De manière analogue, une dépendance entre p_i et la précédente prédiction p_{i-1} peut être considérée. Cela doit notamment permettre d'éviter des enchaînements de phonèmes possiblement non articulables. Nous avons cherché à savoir si cette dépendance est utile.
4. Enfin, il est possible d'interdire ou autoriser la propagation des dépendances par-delà les frontières de mots. Nous avons également étudié ces deux options.

Les CAC permettent naturellement de modéliser tous ces différents types de dépendances (Lafferty *et al.*, 2001) et nous avons pour cela utilisé l'outil Wapiti (Lavergne *et al.*, 2010). Dans le détail, nous avons tout d'abord étudié chaque piste individuellement, puis les avons combinées lors des expériences finales. Avant d'aborder ces travaux, la prochaine section décrit le protocole expérimental.

Les CAC sont entraînés et évalués indépendamment pour chaque locuteur avec l'objectif de déterminer une unique configuration d'apprentissage pour pouvoir généraliser la méthode à tout locuteur. Chaque entretien a été partitionné en ensembles d'entraînement (60 % des énoncés), de développement (20 %) et de test (20 %). L'étude préliminaire des différents paramètres introduits à la section 3 s'est effectuée sur l'ensemble de développement et les expériences finales sur l'ensemble de test.

Pour chaque locuteur, un taux d'erreurs sur les phonèmes (PER, pour *Phoneme Error Rate*) est calculé comme une distance d'édition entre les séquences réalisées et celles canoniques (c.-à-d. non adaptées) ou prédites par l'adaptation. Les résultats reportés dans cet article sont les PER moyens sur la totalité des locuteurs. En complément de ces mesures objectives, des tests d'écoute ont été menés pour mesurer la spontanéité et l'intelligibilité des différentes prononciations étudiées. Ces tests, conduits sur les configurations les plus intéressantes, sont détaillés en section 5.

4 Étude préliminaire des paramètres

Cette section détaille comment la méthode d'adaptation de la prononciation a été réglée sur l'ensemble de développement. L'accent est mis sur les deux aspects majeurs que sont le choix des caractéristiques

	Pas d'adaptation	Phon. canoniques	Car. sélectionnées	Toutes les car.
Unigrammes	30,4	30,4 (0,0)	**24,7 (-5,7)**	26,0 (-4,4)
Uni+bigrammes		25,7 (-4,7)	**24,1 (-6,3)**	26,1 (-4,3)

TABLE 3 – PER (%) sur l'ensemble de développement sans adaptation ou avec adaptation avec différents jeux de caractéristiques.

linguistiques pertinentes et celui de la taille du voisinage à considérer. Néanmoins, l'utilité d'inclure des dépendances entre phonèmes prédits et celle d'adapter à l'échelle de mots isolés ou d'énoncés continus sont également examinées. Le détail de ces études peut être trouvé dans (Qader *et al.*, 2015).

L'entraînement de CAC à partir de trop nombreuses caractéristiques peut conduire à du surapprentissage. Par ailleurs, le temps de calcul et la quantité de mémoire nécessaires à l'entraînement est exponentiel en fonction du nombre de caractéristiques. Ainsi, nous avons effectué une sélection des attributs linguistiques à considérer. Cette sélection s'est faite en recherchant le meilleur ensemble de caractéristiques, c.-à-d. celui qui conduit au plus petit PER, pour chaque locuteur. Nous avons pour cela mis en place un mécanisme de vote. Une caractéristique a reçu un vote par nombre de fois où elle appartenait au meilleur ensemble d'un locuteur. Pour rendre ce processus robuste, deux stratégies de recherche ont été testées pour chaque locuteur : l'une additive, l'autre soustractive. À l'issue des votes, il a arbitrairement été décidé de sélectionner les caractéristiques qui avait reçu au moins $50\,\%$ des votes, soit 20 votes ici. La table 2 reporte entre parenthèses le nombre de votes reçus par chaque caractéristique et en gras celles qui ont finalement été sélectionnées ainsi. Il apparaît que les informations relatives aux syllabes et aux fréquences des mots sont les plus importantes. Ces conclusions sont cohérentes avec de précédents travaux (Adda-Decker *et al.*, 2005; Vazirnezhad *et al.*, 2009; Bell *et al.*, 2009). La table 3 compare les PER obtenus (i) sans adaptation et (ii) avec adaptation sur la base d'aucune caractéristique autre que les phonèmes canoniques, (iii) des caractéristiques sélectionnées et (iv) de toutes les caractéristiques possibles. Ces configurations ont été testées lorsqu'un phonème est prédit soit indépendemment de la prédiction qui le précède (unigrammes), soit en en tenant compte (uni+bigrammes). Il ressort clairement des résultats que la sélection des caractéristiques est une nécessité. Par ailleurs, les dépendances entre prédictions semblent bénéfiques.

Ensuite, comme évoqué en section 3, il peut s'avérer judicieux de considérer les informations provenant du voisinage d'un phonème canonique en cours d'adaptation. Pour déterminer cela, nous avons défini le voisinage comme une fenêtre symétrique [1] de W phonèmes canoniques à gauche et à droite autour du phonème à adapter. Une fenêtre $W{=}0$ signifie ainsi qu'aucun voisinage n'est considéré et $W{=}{\pm}2$ que 5 phonèmes sont considérés au total (1 au centre, 2 à gauche et 2 à droite). La figure 2 présente les PER obtenus pour différentes valeurs de W. Ces résultats sont présentés soit dans le cas où les fenêtres ne peuvent pas traverser des frontières de mots (mots isolés), soit dans celui où elles le peuvent (énoncés). Les CAC ont été appris à partir des seuls phonèmes canoniques et dans la configuration unigramme (pas de dépendances entre phonèmes prédits). Il apparaît que la prise en compte du voisinage améliore significativement les résultats mais qu'un plateau est vite atteint lorsque W augmente. Ainsi, la valeur $W{=}{\pm}2$ est retenue pour les expériences finales. Par ailleurs, contrairement à l'intuition, il semblerait que les adaptations mot à mot produisent de meilleurs résultats que lorsque l'adaptation se fait à l'échelle d'un énoncé entier. L'explication de ce phénomène est que les frontières de mots portaient une information utile quant à la position d'un phonème canonique dans son mot. L'utilisation conjointe de fenêtre et d'énoncés supprime cette information. Nous étayerons cette conclusion grâce aux expériences finales.

1. Nous avons également testé des fenêtre dissymétriques mais celles-ci n'ont mené qu'à de moins bons résultats.

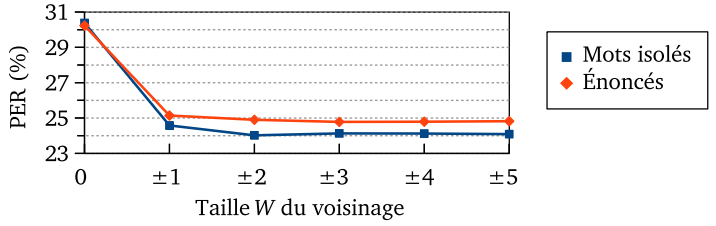


FIGURE 2 – PER en fonction de la taille de la fenêtre, pour les mots isolés et les phrases.

5 Combinaison des paramètres

Cette section présente les expériences finales conduites sur l'ensemble de test de chaque locuteur. Les modèles d'adaptation sont appris en combinant différentes configurations étudiées à la section 4, à savoir :

— à partir des seuls phonèmes canoniques ou en incluant aussi les caractéristiques sélectionnées ;
— sans ou avec prise en compte du voisinage ($W{=}0$ ou $W{=}{\pm}2$) ;
— en incluant ou non une dépendance entre phonèmes prédits (unigrammes ou uni+bigrammes) ;
— en considérant des successions de mots isolés ou, au contraire, des énoncés continus.

La table 4 présente les résultats pour toutes les combinaisons possibles. Les configurations déjà évaluées sur l'ensemble de développement mènent aux mêmes conclusions, à savoir que les prises en compte séparées des caractéristiques linguistiques ou des voisinages produisent des taux d'erreurs plus bas. En outre, ces nouveaux résultats montrent que le croisement de ces deux configurations permet d'obtenir d'encore meilleurs PER. Ces différentes variations sont, certes, faibles mais elles sont statistiquement significatives avec un niveau de confiance de 95 %. Ensuite, les résultats montrent que l'adaptation d'énoncés continus produit de meilleurs résultats absolus que sur des suites de mots isolés. Cet écart, déjà observé sans adaptation, est en réalité dû au fait que des erreurs successives d'insertion et de suppression à la frontière de mots sont considérées comme une seule substitution pour les énoncés. Ces résultats sur les énoncés nous permettent néanmoins de noter que l'inclusion des caractéristiques linguistiques permet de corriger le décalage précédemment observé sur la figure 2 par la réintroduction d'informations sur la position d'un phonème dans son mot. Enfin, nous pouvons noter que la prise en compte des dépendances entre phonèmes prédits devient néfaste lorsque beaucoup de caractéristiques sont prises en compte conjointement. Nous expliquons ces résultats par un phénomène de surapprentissage lié à l'explosion du nombre de paramètres à estimer par le CAC. Pour illustrer ces résultats, la table 5 compare les prononciations réalisées, canoniques et obtenues par deux adaptations sur l'exemple déjà présenté à la table 1. Celui-ci montre que l'adaptation sur la base des caractéristiques linguistiques et des voisinages introduit des variantes de prononciations plus relâchées. Il montre néanmoins également que l'absence de dépendances entre les phonèmes prédits peut conduire à des séquences peu plausibles, comme ici l'enchaînement /n̬n/.

Pour vérifier cette explication et approfondir l'intérêt de tenir compte des dépendances entre phonèmes prédits, nous avons conduit une autre série d'expériences. Nous avons appris un modèle de langage (ML) sur les phonèmes réalisés de l'ensemble des données d'apprentissage [2], puis avons utilisé ce modèle pour réordonner *a posteriori* les meilleures hypothèses d'adaptation fournies par nos CAC. Précisément, chaque hypothèse h est associée à un score $s(h)$ calculé comme une

2. Un seul ML pour les 20 locuteurs a été appris plutôt qu'un par locuteur afin d'estimer des probabilités fiables.

			$W = 0$	$W = \pm 2$
Suites de mots isolés				
Phonèmes canoniques (c.-à-d. pas d'adaptation)			30,5	–
Phonèmes adaptés à partir des	phonèmes canoniques seuls	Unigrammes	30,4 (-0,1)	23,8 (-6,7)
		Uni+bigrammes	25,5 (-5,0)	24,0 (-6,5)
	+ caractéristiques linguistiques	Unigrammes	24,3 (-6,2)	**23,6 (-6,9)**
		Uni+bigrammes	24,1 (-6,4)	24,2 (-6,3)
Énoncés continus				
Phonèmes canoniques (c.-à-d. pas d'adaptation)			30,3	–
Phonèmes adaptés à partir des	phonèmes canoniques seuls	Unigrammes	30,2 (-0,1)	24,9 (-5,4)
		Uni+bigrammes	25,9 (-4,4)	24,2 (-6,1)
	+ caractéristiques linguistiques	Unigrammes	24,1 (-6,2)	**23,4 (-6,9)**
		Uni+bigrammes	23,9 (-6,4)	24,4 (-5,9)

TABLE 4 – PER (%) sur l'ensemble de test. Entre parenthèses, les variations absolues avec les PER des prononciations canoniques seuls (sans adaptation).

Phonèmes réalisés		/k ɑ n s n̩ _ t ɹ eɪ _ ɪ d · ɪ f̃ · oʊ h ɑ ʌ /
Phonèmes canoniques (c.-à-d. pas d'adaptation)		/k ɑ n s **ʌ n** t ɹ eɪ **t ʌ** d · ɪ **n** · oʊ h **ɑɪ oʊ** / (7 erreurs)
Phonèmes adaptés à partir des	phonèmes canoniques seuls	/k ɑ n s **ʌ n** t ɹ eɪ **t ʌ** d · ɪ **n** · oʊ h **ɑɪ oʊ** / (7 erreurs)
	+ carac. ling. + voisinage	/k ɑ n s n̩ **n̩** ɹ eɪ **ɾ** ɪ d · ɪ **n** · oʊ h **ɑɪ oʊ** / (6 erreurs)

TABLE 5 – Différentes prononciations de la séquence de mots « *concentrated in Ohio* ». Les prononciations adaptées ont été produites sur la base de mots isolés sans prise en compte des dépendances entre phonèmes prédits. Les erreurs par rapport à la référence sont reportées en gras.

interpolation logarithmique des probabilités fournies par le CAC et par le ML, comme suit :

$$s(h) = \Pr{}_{\text{CAC}}(h) \times \Pr{}_{\text{ML}}(h)^{\alpha} \times \beta^{n} \,, \tag{1}$$

où α et β sont deux paramètres à optimiser et n est le nombre de phonèmes dans h. Le facteur β sert à contrebalancer le favoritisme naturel du ML envers les hypothèses les plus courtes. Le réordonnancement consiste alors à sélectionner l'hypothèse de score s le plus élevé. En pratique, le ML est un modèle 5-gramme avec un lissage de Witten-Bell et α et β ont été optimisés de sorte à minimiser le PER sur l'ensemble de développement. L'apprentissage du ML, l'optimisation des paramètres et le réordonnancement ont été effectués grâce à l'outil SRILM (Stolcke *et al.*, 2011). En reprenant les meilleurs résultats de la table 4 (sur la configuration « unigrammes »), l'introduction par le ML des dépendances entre phonèmes prédits permet d'obtenir des améliorations additionnelles significatives du PER sur l'ensemble de test, respectivement de $0,3$ et $0,2$ point pour les mots isolés et les énoncés continus. Ces résultats confortent en outre notre hypothèse sur l'effet négatif d'un trop grand nombre de paramètres lors de l'apprentissage des CAC.

La spontanéité et l'intelligibilité a été évaluée perceptuellement par un test d'écoute AB sur 10 locuteurs anglais natifs. Ce test juge la préférence des testeurs entre des paires de signaux de parole synthésisés sur la base des prononciations non adaptées, adaptées à partir des phonèmes canoniques (C) ou également des informations linguistiques (C+L), ou réalisées. Toutes ces configurations inclus le réordonnancement des hypothèses. Le test contient 40 étapes [3]. À chaque étape, le testeur évalue sa

3. Quelques échantillons peuvent être écoutés sur `http://www-expression.irisa.fr/demos`.

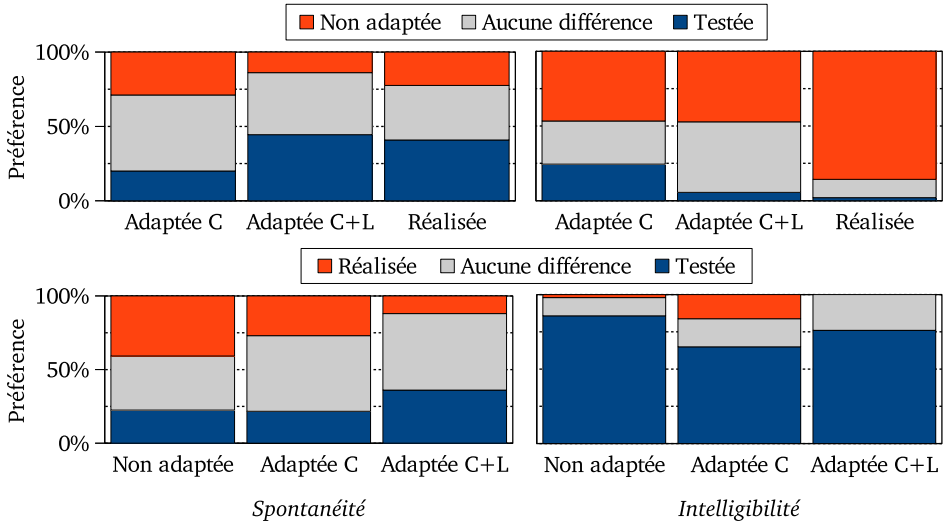


FIGURE 3 – Comparaisons entre les prononciations adaptées et les prononciations non adaptées (en haut) ou réalisées (en bas) en termes de spontanéité et d'intelligibilité.

préférence entre les 2 échantillons proposés en terme de spontanéité et d'intelligibilité. Le système est système HTS classique appris sur les données du challenge Blizzard 2012. Les résultats de ce test sont présentés par la figure 3. Il en ressort que les prononciations adaptées et réalisées sont effectivement jugées plus spontanées que les prononciations non adaptées. Tout particulièrement, il apparaît même que la prise en compte des informations linguistiques pour l'adaptation aboutit à des prononciations qui, une fois synthétisées, sont jugées plus spontanées que les prononciations réalisées. Enfin, les prononciations non adaptées sont les plus intelligibles mais les prononciations adaptées sont, là encore, bien meilleures sur ce point que les prononciations réalisées, notamment lorsque les informations linguistiques sont considérées.

6 Conclusion et perspectives

Dans cet article, nous avons proposé une méthode d'adaptation de la prononciation qui permet d'imiter un style spontané. Les prononciations adaptées sont destinées à améliorer les systèmes de synthèse de la parole. Appliquées à l'anglais, les expériences sur le corpus de parole Buckeye montrent d'ores et déjà que les prononciations adaptées reflètent significativement mieux les prononciations spontanées des locuteurs que les prononciations canoniques d'origine. Ces bons résultats sont en particulier atteints grâce à la prise en compte de caractéristiques linguistiques sélectionnées automatiquement, de voisinages autour de chaque phonème canonique et de dépendances entre phonèmes adaptés.

Parmi les perspectives, ce travail pourrait être complété par une étude sur la prise en compte de caractéristiques articulatoires, prosodiques et acoustiques. Notre méthode pourrait en outre être testée sur d'autres langues ou sur des prononciations fournies par des phonétiseurs automatiques. Ces perspectives sont actuellement en cours d'étude. Notons enfin que, à terme, notre travail pourrait également trouver des applications en reconnaissance automatique de la parole.

Références

ADDA-DECKER M., DE MAREÜIL P. B., ADDA G. & LAMEL L. (2005). Investigating syllabic structures and their variation in spontaneous French. *Speech Communication*, **46**(2).

BATES R. & OSTENDORF M. (2002). Modeling pronunciation variation in conversational speech using prosody. In *ISCA Tutorial and Research Workshop (ITRW) on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*.

BELL A., BRENIER J. M., GREGORY M., GIRAND C. & JURAFSKY D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, **60**(1).

BELL A., JURAFSKY D., FOSLER-LUSSIER E., GIRAND C., GREGORY M. & GILDEA D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, **113**(2).

CHEN K. & HASEGAWA-JOHNSON M. (2004). Modeling pronunciation variation using artificial neural networks for English spontaneous speech. In *Proc. of Interspeech*.

DILTS P. C. (2013). *Modelling phonetic reduction in a corpus of spoken English using random forests and mixed-effects regression*. PhD thesis, University of Alberta.

FOSLER-LUSSIER E. *et al.* (1999). Multi-level decision trees for static and dynamic pronunciation models. In *Proc. of Eurospeech*.

ILLINA I., FOHR D. & JOUVET D. (2011). Grapheme-to-phoneme conversion using conditional random fields. In *Proc. of Interspeech*.

KARANASOU P., YVON F., LAVERGNE T. & LAMEL L. (2013). Discriminative training of a phoneme confusion model for a dynamic lexicon in ASR. In *Proc. of Interspeech*.

KOLLURU B., WAN V., LATORRE J., YANAGISAWA K. & GALES M. J. F. (2014). Generating multiple-accent pronunciations for TTS using joint sequence model interpolation. In *Proc. of Interspeech*.

LAFFERTY J., MCCALLUM A. & PEREIRA F. C. (2001). Conditional random fields : probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proc. of ACL*.

LECORVÉ G. & LOLIVE D. (2015). Adaptive statistical utterance phonetization for French. In *Proc. of ICASSP*.

PITT M. A., JOHNSON K., HUME E., KIESLING S. & RAYMOND W. (2005). The Buckeye corpus of conversational speech : labeling conventions and a test of transcriber reliability. *Speech Communication*, **45**(1).

PRAHALLAD K., BLACK A. W. & MOSUR R. (2006). Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis. In *Proc. of ICASSP*, volume 1.

QADER R., LECORVÉ G., LOLIVE D. & SÉBILLOT P. (2015). Probabilistic Speaker Pronunciation Adaptation for Spontaneous Speech Synthesis Using Linguistic Features. In *Proc. of SLSP*.

STOLCKE A., ZHENG J., WANG W. & ABRASH V. (2011). Srilm at sixteen : Update and outlook. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, p.5̃.

VAZIRNEZHAD B., ALMASGANJ F. & AHADI S. M. (2009). Hybrid statistical pronunciation models designed to be trained by a medium-size corpus. *Computer Speech & Language*, **23**(1).

WANG D. & KING S. (2011). Letter-to-sound pronunciation prediction using conditional random fields. *IEEE Signal Processing Letters*, **18**(2).

Alignement de séquences phonétiques pour une analyse phonologique des erreurs de transcription automatique

Camille Dutrey [1,2] Martine Adda-Decker [1,3] Naomi Yamaguchi [1]
(1) Laboratoire de Phonétique et Phonologie (LPP), 19 rue des Bernardins, Paris, France
(2) Laboratoire National de Métrologie et d'Essais (LNE), 29 avenue Roger Hennequin, Trappes, France
(3) Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), Rue John Von Neumann, Orsay, France
camille.dutrey@lne.fr, {madda, naomi.yamaguchi}@univ-paris3.fr

RÉSUMÉ

La transcription automatique de la parole obtient aujourd'hui des performances élevées avec des taux d'erreur qui tombent facilement en dessous de 10% pour une parole journalistique. Cependant, pour des conversations plus libres, ils stagnent souvent autour de 20–30%. En français, une grande partie des erreurs sont dues à des confusions entre homophones n'impliquant pas les niveaux acoustico-phonétique et phonologique. Cependant, de nombreuses erreurs peuvent s'expliquer par des variantes de productions non prévues par le système. Afin de mieux comprendre quels processus phonologiques pourraient expliquer ces variantes spécifiques de la parole spontanée, nous proposons une analyse des erreurs en comparant prononciations attendue (référence) et reconnue (hypothèse) *via* un alignement phonétique par programmation dynamique. Les distances locales entre paires de phonèmes appariés correspondent au nombre de traits phonétiques disjoints. Nos analyses permettent d'identifier les traits phonétiques les plus fréquemment impliqués dans les erreurs et donnent des pistes pour des interprétations phonologiques.

ABSTRACT

Phonetic sequences alignment for a phonemic analysis of automatic speech transcription errors

Nowadays, word error rates of automatic speech transcription systems tend to fall below 10% for journalistic speech. However, in the case of free conversations, error rates remain much higher, typically around 20-30%. Error sources range from system limitations such as out of vocabulary words to speaker production errors. In French, many errors are due to homophonic words, for which neither acoustic-phonetic nor phonological levels are to blame. An important part may be related to production variants unknown to the system. To investigate which phonological processes might contribute to explain fluent speech specific variants, a phone sequence alignment between reference and hypothesis phone strings was implemented using dynamic programming. Local distances are computed as the total number of disagreeing phonetic features between phone pairs. The resulting analyses highlight the features most frequently involved in recognition errors and provide insight for phonological interpretations of fluent speech variation.

MOTS-CLÉS : alignement de séquences, traits distinctifs, programmation dynamique, reconnaissance automatique de la parole, erreurs de transcription.

KEYWORDS: sequence alignment, distinctive features, dynamic programmation, automatic speech recognition, transcription errors.

1 Introduction

Dans cette contribution, nous proposons d'analyser les erreurs de transcription automatique de la parole d'un point de vue phonétique. Les erreurs d'un système de transcription sont habituellement comptabilisées au niveau du mot et une confusion entre deux formes fléchies homophones (p. ex. « politique » et « politiques ») compte autant qu'une erreur entre mots très différents (p. ex. « affaire » et « ferveur » dans la suite « l'affaire Woerth » reconnue comme « la ferveur »).

Pour cela, nous comparons la chaîne phonétique correspondant aux mots de la transcription automatique (hypothèse ou HYP) à celle provenant des mots de la transcription manuelle (référence ou REF). La comparaison est effectuée dans des zones d'erreur, c'est-à-dire aux endroits où le système de transcription produit des mots différents de ceux attendus par la référence. Pour comparer des chaînes de caractère, une mesure fréquemment utilisée est la distance d'édition ou la distance de Levenshtein (Levenshtein, 1965), qui donne le nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaîne à l'autre. Dans notre cas, les caractères correspondent à des phones. L'alignement de séquences phonétiques a été utilisé pour de nombreuses recherches en phonologie computationnelle (Kondrak, 2000, 2003), en dialectologie (Heeringa, 2004; Heeringa *et al.*, 2002) ou encore en phonétique clinique (Connolly, 1997).

Nous adoptons ici ce type d'approche pour l'analyse d'erreurs issues de systèmes de transcription automatique de la parole. Nous proposons d'adapter la distance de Levenshtein pour mieux tenir compte de la proximité phonético-phonologique entre phonèmes. Par exemple, le /p/ est plus proche du /b/ que du /s/ ou du /a/.

2 Corpus, approche et méthode

Nos travaux s'inscrivent dans le cadre plus large de recherches menées sur la caractérisation des erreurs produites par des systèmes de transcription de la parole, notamment au sein du projet ANR VERA [1] (Goryainova *et al.*, 2014; Luzzati *et al.*, 2014; Santiago *et al.*, 2015). L'objectif est d'étudier l'impact de ces erreurs sur des applications plus complexes comme l'indexation, la traduction ou le repérage d'entités nommées à partir de flux audio. D'autres finalités consistent à contribuer à une évaluation plus informative des systèmes de transcription et à mieux rendre compte des aspects linguistiques impliqués dans les erreurs. Nous développons ce dernier volet en focalisant sur l'interface phonétique–phonologie dans les erreurs de transcription des sorties du système de reconnaissance du LIUM (Bougares *et al.*, 2013) avec les données de la campagne ETAPE (Gravier *et al.*, 2012). Dans la suite, nous présentons d'abord le corpus qui fournit les erreurs de transcription, c'est-à-dire les séquences phonétiques (REF *vs* HYP) à aligner. Nous développons ensuite le choix de traits phonétiques pour décrire les phonèmes du français, utilisés pour le calcul de distance entre paires de phonèmes. Nous précisons que ce calcul de distance se fait uniquement à partir de traits sans prendre en compte les réalisations acoustiques des sons impliqués. Enfin, nous rappelons brièvement l'algorithme de programmation dynamique tel que mis en œuvre pour notre analyse.

1. http://projet-vera.univ-lemans.fr/.

2.1 Corpus de parole préparée et spontanée : ETAPE

Nous avons travaillé sur un corpus de parole journalistique préparée et spontanée constitué d'émissions radio et télé-diffusées, le corpus ETAPE (Gravier *et al.*, 2012). Nous en avons traité 58 enregistrements, ce qui représente environ 35h de parole pour 339k mots prononcés. La transcription automatique produite par le système du LIUM, comporte 323k mots sur ce sous-ensemble d'ETAPE ; pour une description détaillée de ce système et de sa paramétrisation, se référer à Bougares *et al.* (2013). Afin d'étudier les erreurs de transcription automatique d'un point de vue phonétique, deux étapes préliminaires ont été réalisées :

1. un alignement REF *vs* HYP au niveau des mots à l'aide du NIST Scoring Toolkit [2] pour obtenir les types d'erreur (correct *versus* substitution, suppression et insertion) ;
2. une phonétisation des mots réalisée avec le système d'alignement forcé du LIMSI (Gauvain *et al.*, 2003) et un jeu de 33 phonèmes du français : à noter l'absence du /œ/ supplanté par /ə/. L'alignement est réalisé sur l'ensemble de l'audio, à la fois pour la référence et l'hypothèse, en utilisant le même dictionnaire de prononciation (dictionnaire du LIUM) que celui utilisé par le système de reconnaissance de la parole.

REF	« donc	le	**fort**	**taux**	**de**	natalité »
HYP	« donc	le	*	*	**forte**	natalité »
erreur	C	C	D	D	S	C

FIGURE 1 – Extrait de parole transcrite avec alignement REF *versus* HYP et indication du type d'erreur assignée par le système pour chaque mot (C = correct ; D = suppression ; S = substitution).

Nous avons extrait du corpus 18 051 zones d'erreurs : une zone d'erreur correspond à une suite ininterrompue de mots erronés entre deux mots non erronés. La figure 1 donne un exemple d'énoncé où la zone d'erreur est marquée en gras. Les données ont été pré-traitées de manière à en exclure les zones d'erreur trop complexes résultant souvent de décalages entre REF et HYP. Cette sélection, qui exclut 13,6 % des zones d'erreur, s'appuie sur des critères liés à la différence de longueur entre la séquence phonétique de REF et celle de HYP. Nous avons également choisi d'écarter les zones d'erreur incluant des phénomènes particuliers de la parole spontanée, qui seront analysés à part : hésitations vocaliques, présence d'amorces de mot, *etc.* Au final, 16,5 % des zones d'erreur ont ainsi été mises de côté et 13 021 zones d'erreurs sont conservées.

2.2 Spécification en traits des phonèmes du français

La comparaison des phonèmes du français s'appuie sur la construction d'une matrice de traits distinctifs : les traits pris en compte sont uniquement les traits distinctifs en français et sont adaptés de Sagey (1986) et Walker (1993). Ces 13 traits ont été considérés comme privatifs (Mester & Ito, 1989) pour la spécification des phonèmes du français. La spécification est par ailleurs totale : tous les phonèmes sont spécifiés pour tous les traits, y compris lorsque la valeur de trait est redondante (p. ex. [voisé] pour les sonantes), comme présenté dans le tableau 1. Dans cette perspective, nous parlerons alors de traits phonétiques. Dans cette étude, le terme « phonème » englobe consonnes, voyelles et semi-voyelles, même si ces dernières peuvent être considérées comme des allophones de certaines voyelles.

2. Outil accessible à l'adresse Web suivante : http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sctk.htm.

/p/	cons., labial
/b/	cons., labial, voisé
/t/	cons., coronal
/d/	cons., coronal, voisé
/k/	cons., dorsal
/g/	cons., dorsal, voisé
/f/	cons., cont., labial
/v/	cons., cont., labial, voisé
/s/	cons., cont., coronal
/z/	cons., cont., coronal, voisé
/ʃ/	cons., cont., coronal, post.
/ʒ/	cons., cont., coronal, post., voisé
/l/	cons., cont., coronal, sonant, voisé, latéral
/m/	cons., cont., labial, sonant, voisé, nasal
/n/	cons., cont., coronal, sonant, voisé, nasal
/ɲ/	cons., cont., coronal, sonant, voisé, nasal, post.
/ʁ/	cons., cont., dorsal, sonant, voisé

(a) Consonnes.

/i/	cont., coronal, sonant, voisé, haut
/y/	cont., coronal, sonant, voisé, haut, arrondi
/u/	cont., dorsal, sonant, voisé, haut, arrondi
/e/	cont., coronal, sonant, voisé
/ø/	cont., coronal, sonant, voisé, arrondi
/o/	cont., dorsal, sonant, voisé, arrondi
/ɛ/	cont., coronal, sonant, voisé, bas
/ə/	cont., sonant, voisé
/ɔ/	cont., dorsal, sonant, voisé, bas, arrondi
/a/	cont., dorsal, sonant, voisé, bas
/ɔ̃/	cont., dorsal, sonant, voisé, bas, arrondi, nasal
/ɛ̃/	cont., sonant, voisé, bas, nasal
/ɑ̃/	cont., dorsal, sonant, voisé, bas, nasal
/j/	cont., coronal, sonant, voisé, haut
/w/	cont., dorsal, sonant, voisé, haut, arrondi
/ɥ/	cont., coronal, sonant, voisé, haut, arrondi

(b) Voyelles et semi-voyelles.

TABLE 1 – Spécification en traits pour les phonèmes du français utilisée pour le calcul de distance et l'alignement de séquences phonétiques (cons. = consonantique ; cont. = continu ; post. = postérieur).

Le trait [consonantique] distingue les consonnes des voyelles et semi-voyelles, et représente le degré de constriction dans le conduit vocal. Les traits de lieu [labial], [coronal] et [dorsal] indiquent le lieu d'articulation des sons ; le trait [postérieur] distingue dans les consonnes [coronal] les sons produits avec l'arrière de la langue de ceux produits avec l'avant de la langue. Le trait [voisé] désigne le voisement des phonèmes. Le trait de mode [sonant] distingue les sonantes des obstruantes ; [continu] indique le passage continu de l'air dans le conduit vocal, et distingue les sonantes et fricatives des occlusives. [nasal] est spécifié pour les sons laissant passer l'air par la cavité nasale. Les traits vocaliques [haut] et [bas] caractérisent l'aperture. Le trait [arrondi] spécifie les (semi-)voyelles produites avec un arrondissement des lèvres. Chaque phonème est ainsi représenté par un vecteur de dimension 13 (V_φ) avec des 0 pour tous les traits non-spécifiés et des 1 pour les traits spécifiés.

2.3 Calcul de distances pour la comparaison de paires de phonèmes

Nous avons calculé, pour chaque paire de phonèmes (φ_i, φ_j) du français [3], une distance phonétique $d(i, j)$ (*cf.* section 2.4) à partir de la spécification en traits présentée ci-dessus. La distance $d(i, j)$ correspond à la somme de l'opérateur ou-exclusif sur les 13 dimensions des vecteurs correspondant aux phonèmes φ_i et φ_j ($V_{\varphi_i} \oplus V_{\varphi_j}$) et donne le nombre de traits phonétiques disjoints. Alors que théoriquement la valeur maximale pourrait être 13 pour un vecteur de dimension 13, il se trouve que les traits sont distribués de telle manière que la distance maximale se limite à 9. La distribution de ces distances est présentée en figure 2, selon le type de paire : voyelle *versus* voyelle ; consonne *versus* consonne ; voyelle *versus* consonne. 36 paires ont une distance nulle : il s'agit de comparaisons de phonèmes identiques, en tenant compte des allophones. Ainsi, les paires /i/-/j/, /y/-/ɥ/ et /u/-/w/ sont également distantes de 0.

3. Soit 561 combinaisons ; le nombre de paires correspond au nombre de cases d'une matrice triangulaire hors diagonale ($33 \times 32/2$) plus la diagonale (33).

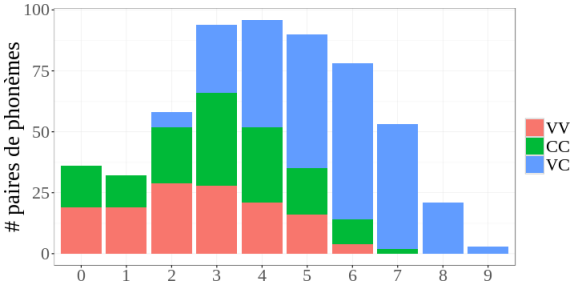


FIGURE 2 – Distribution des paires de phonèmes en fonction de leur distance locale, avec caractérisation du type de paire comparée (VV = voyelles ; CC = consonnes ; VC = voyelle *vs* consonne).

On peut remarquer que les distances faibles correspondent majoritairement à des paires "homogènes" VV et CC et que la proportion de paires CV augmente au fur et à mesure que la distance augmente.

Le tableau 2 recense les paires de phonèmes appartenant aux distances maximales et minimales (hors distances nulles) pour chaque type de paires (VV, CC et VC). On remarque la présence de /p/ et /t/ dans les paires de CV maximalement distantes : ceci peut s'expliquer par le fait que ces phonèmes, particulièrement /t/, sont considérés comme des sons phonologiquement non marqués (Paradis & Prunet, 1991) et sont de ce fait spécifiés par moins de traits que les autres phonèmes. Les distances entre phonèmes sont utilisées dans cette étude comme connaissance linguistique dans le programme d'alignement de séquences phonétiques par programmation dynamique.

Voyelle-Voyelle	min=1	/i-y/ /i-e/ /i-ɥ/ /y-ø/ /y-j/ /u-o/ /e-ø/ /e-ɛ/ /e-ə/ /e-j/ /ø-ɥ/ /o-ɔ/ /o-w/ /ɔ-a/ /ɔ̃-ɔ̃/ /a-ɑ̃/ /ɔ̃-ɑ̃/ /ɛ̃-ɑ̃/ /j-ɥ/
	max=6	/i-ɔ̃/ /y-ɑ̃/ /ɔ̃-j/ /ɑ̃-ɥ/
Consonne-Consonne	min=1	/p-b/ /p-f/ /b-v/ /t-d/ /t-s/ /d-z/ /k-g/ /f-v/ /s-z/ /s-ʃ/ /s-/ /ʃ-ʒ/ /n-ɲ/
	max=7	/p-ɲ/ /k-ɲ/
Voyelle-Consonne	min=2	/e-z/ /e-l/ /e-n/ /o-ʁ/ /ə-ʁ/ /a-ʁ/
	max=9	/ɔ̃-p/ /ɔ̃-t/ /ɔ̃-ʃ/

TABLE 2 – Paires de phonèmes impliquées dans les distances minimales et maximales par type (VV = paires de voyelles ; CC = paires de consonnes ; VC = paires de voyelle *versus* consonne).

2.4 Mesure de distances phonétiques et alignement de séquences

Le programme d'alignement de séquences phonétiques est adapté de l'algorithme de Levenshtein (1965) qui permet de calculer des mesures de distances entre deux chaînes de caractères. Il s'appuie sur la programmation dynamique (Bellman, 1957; Vintsyuk, 1968), dont le principe est rapidement décrit ci-dessous. Soient deux séquences phonétiques $\Phi_I = \varphi_1\varphi_2 \dots \varphi_I$ (hypothèse) et $\Phi_J = \varphi_1\varphi_2 \dots \varphi_J$ (référence) de longueur I et J respectivement. On impose le départ de l'alignement au début des deux chaînes respectives ce qui se traduit par des conditions d'initialisation $D(0,0) = 0$, $D(0,j) = \infty$ pour $j > 0$, $D(i,0) = \infty$ pour $i > 0$. Ensuite la récurrence sur i, j (chaînes partielles de 1 à i et de 1 à j) s'écrit comme suit :

$$D(i,j) = min \begin{cases} D(i-1,j) + d(i,j) & \text{insertion} \\ D(i,j-1) + d(i,j) & \text{omission} \\ D(i-1,j-1) + 2 \times d(i,j) & \text{correct ou substitution} \end{cases} \quad (1)$$

où $d(i,j) = 0$ si $\varphi_i = \varphi_j$. L'arrêt se fait naturellement à (I, J), à la fin des deux chaînes. Pour récupérer l'alignement correspondant à la distance globale minimale, nous avons introduit dans la récurrence une matrice de *retour-arrière*, qui à chaque point (i, j) garde la mémoire du meilleur point précédent (argmin de l'équation 1). La distance globale, qui permet de caractériser la dissimilarité phonétique entre deux séquences de phonèmes, est ensuite normalisée par le nombre de phonèmes dans la référence. Elle nous permet d'analyser plus finement les erreurs commises par les systèmes de transcription de la parole, notamment grâce à son appui sur des connaissances phonologiques.

Dans un premier temps, nous souhaitons limiter nos analyses à des zones d'erreurs que nous jugeons intéressantes d'un point de vue phonétique comme phonologique. Dans ce but, nous rejetons dans la suite celles dont les longueurs sont très différentes entre REF et HYP et pour lesquelles des problèmes de découpage du signal ou de bruit de fond viennent supplanter les facteurs linguistiques. Nous gardons ainsi 11 753 zones d'erreur des données ETAPE, dont les distances globales normalisées se distribuent entre 0 et 5 (*cf.* figure 4). Pour ces zones d'erreur, l'information de *retour-arrière* a été utilisée pour récupérer l'alignement de la séquence phonème à phonème. En effet, l'alignement des zones d'erreur au niveau du phonème, comme illustré en figure 3, permet de mieux décrire et analyser les erreurs produites par le système de transcription de la parole d'un point de vue phonétique. Comme le met en exergue l'exemple illustré, cet alignement permet d'obtenir une finesse de localisation et d'identification phonétique des erreurs totalement absente des méthodes classiques qui évaluent les systèmes de reconnaissance automatique de la parole au seul niveau de la séquence de mots.

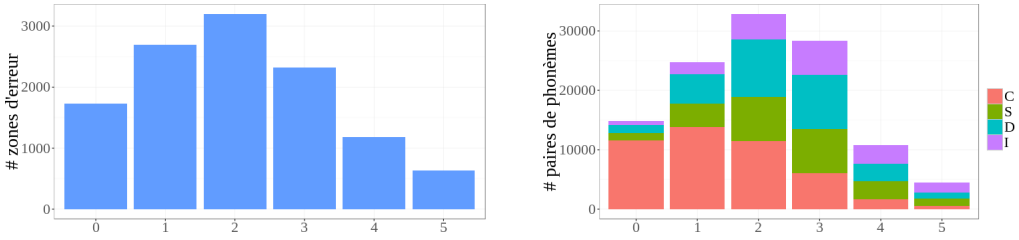
REF	**fort**	**taux**	**de**	⇒	[	f	ɔ	ʁ	t	**o**	**d**	]
HYP			**forte**	⇒	[	f	ɔ	ʁ	t	**ə**	*****	]
erreur	D	D	S			C	C	C	C	S	D	

FIGURE 3 – Extrait de parole transcrite avec comparaison de l'alignement REF *versus* HYP produit par un système d'évaluation de la transcription automatique (au niveau du mot) et de celui produit par le système d'alignement de séquences phonétiques (C = correct ; D = suppression ; S = substitution).

La figure 3 illustre une zone d'erreur de distance globale normalisée égale à 1 pour laquelle l'évaluation classique produit deux omissions de mots et une substitution et qui, d'un point de vue phonétique, se révèle majoritairement correcte. L'erreur commise peut s'expliquer d'un côté par une forte réduction temporelle de l'article « de » dans le contexte « fort taux de natalité » : le schwa est tombé et le [d] se limite à environ 30 ms de barre de voisement avant le [n]. Dans ce contexte de quasi-absence du *de*, le modèle de langage impose « forte natalité » plutôt que « fort taux de natalité ».

3 Analyse des erreurs de transcription automatique

La méthode présentée dans cette étude, permettant de calculer des distances entre chaînes phonétiques au-delà des mots et d'aligner ces dernières en tenant compte d'informations phonologiques, peut être mise au service d'une analyse linguistique des erreurs de transcription de la parole. Nous souhaitons ainsi contribuer à l'évaluation des systèmes de transcription de la parole en utilisant ces informations de manière à mieux identifier les variations impactant les phonèmes et le rôle des traits phonétiques.



(a) distribution des zones d'erreur.

(b) distribution des paires de phonèmes.

FIGURE 4 – Caractérisation des zones d'erreur : (a) distribution des zones d'erreur en fonction de la distance globale normalisée entre REF et HYP ; (b) distribution des paires de phonèmes impliquées par type d'erreur subie en fonction de la distance globale normalisée (C = correct ; S = substitution ; D = suppression ; I = insertion).

3.1 Caractérisation de zones d'erreurs par distance phonétique

La figure 4 permet de visualiser la distribution des zones d'erreur en fonction de leur distance phonétique. Une analyse préliminaire qualitative permet de faire l'hypothèse que cette mesure de distance pourrait permettre de catégoriser efficacement les zones d'erreur produites par les systèmes de transcription automatique. En effet, les zones présentant une distance normalisée nulle permettent d'identifier des chaînes homophoniques ou quasi-homophoniques, telles « leaders » *versus* « lits de leur » ou « base » *versus* « basse », y compris sur des séquences relativement longues, comme « vin de Féternes » *versus* « vingt-deux faits termes ».

Les séquences présentant une distance moyenne (p. ex. 2, distance la plus fréquente du corpus) sont phonétiquement proches tout en présentant, notamment, de nombreuses substitutions sur des paires de type VV ou CC, comme dans « que ce label » [ksəlabɛl] *versus* « solennel » [sɔlanɛl] ou des prononciations non-canoniques sources de confusion, comme dans « sans sans langue de bois » [sɑ̃sɑ̃lɑ̃gədəbwa] *versus* « cinq cent emplois » [sɛ̃ksɑ̃ɑ̃plwa] (pour cet exemple, la transcription automatique est également mise en difficulté par de la parole superposée). Enfin, les séquences présentant une distance maximale (p. ex. « bon Copé » *vs* « à la rentrée ») sont souvent particulièrement difficiles à transcrire, avec beaucoup de parole superposée ou de bruit environnant.

3.2 Implication des traits phonétiques dans les zones d'erreurs

L'examen des traits phonétiques impliqués dans les zones d'erreur (*cf.* figure 5) indique que tous les traits n'ont pas la même importance dans les erreurs. Les traits les mieux reconnus sont les traits [continu], [voisé], [sonant], [consonantique] qui sont les traits les plus fréquents, et qui correspondent à des distinctions fondamentales phonologiques (Clements, 1985). Les traits qui sont plus souvent substitués que bien reconnus sont les traits [arrondi] et [postérieur], qui ne distinguent respectivement qu'une petite partie des voyelles et des consonnes. Quant aux suppressions et aux insertions, elles concernent majoritairement les traits qui sont partagés par de nombreux phonèmes : [continu] et [voisé] (spécifiés pour 27 phonèmes), [sonant] (spécifié pour 21 phonèmes). Ces premières observations semblent indiquer que les traits participent dans les différents types d'erreurs en fonction de leur rôle dans le système phonologique et de leur place dans une séquence de sons. Ces résultats sont bien

entendu à approfondir par l'analyse des combinaisons de traits impliqués dans chaque type d'erreur, et par l'étude des séquences de traits dans les suites de phonèmes.

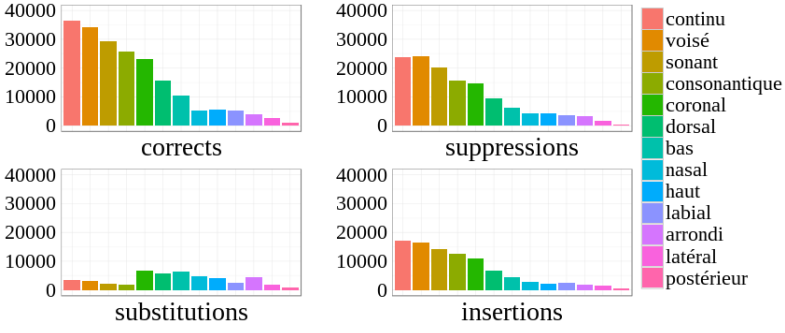


FIGURE 5 – Distribution des traits phonétiques selon leur implication dans des types d'erreurs.

4 Conclusion

Nous avons présenté nos travaux sur l'analyse des erreurs produites d'un système de transcription de l'évaluation ETAPE. Afin d'analyser des zones d'erreur (incluant tous les mots erronés entre les deux extrémités bien reconnues), nous avons introduit une nouvelle méthodologie s'appuyant sur la programmation dynamique en traitement automatique et les traits phonétiques. Cette approche vise à aligner non plus les mots en tant que tels, mais leurs séquences phonétiques respectives afin de mieux décrire les erreurs en matière de proximité phonique. L'utilisation des traits met en lumière le rôle de certains traits, eux-mêmes importants dans le système phonologique, dans les erreurs du système de transcription. En particulier, ce résultat apporte des arguments en faveur de la structuration du système phonologique en traits hiérarchisés (Clements, 1985). Les résultats permettent de trier les zones d'erreur suivant une distance phonétique : à distance faible, nous sommes en présence d'erreurs homophones et quasi-homophones. Au sein de ce sous-ensemble, il sera intéressant d'étudier plus finement les processus phonétiques et phonologiques (lénition, assimilation, chute de segments, réductions et divers metaplasmes). Plusieurs améliorations de la procédure d'alignement sont prévues, et notamment : affiner la représentation des traits phonétiques (p. ex. type de spécification) ; améliorer le calcul des distances locales en tenant compte du voisinage phonétique immédiat. Nous avons également pour perspective de développer les analyses à l'interface phonétique-phonologie, de produire des analyses d'erreur en contexte (analyse des triphones) et fonction des frontières de mots/syllabes et de mieux rendre compte des phénomènes de réduction en parole spontanée. Enfin, nous envisageons dans le futur d'ajouter un décodage phonétique afin d'étudier le rôle des traits phonétiques dans les erreurs de reconnaissance automatique hors contraintes lexicales (et hors contraintes de plus haut niveau).

Remerciements

Ce travail a été partiellement financé par l'Agence Nationale de la Recherche au titre du projet VERA (ANR-12-BS02-006-01) et du programme Investissements d'Avenir (ANR-10-LABX-0083).

Références

BELLMAN R. (1957). *Dynamic Programming*. Princeton University Press.

BOUGARES F., DELÉGLISE P., ESTÈVE Y. & ROUVIER M. (2013). LIUM ASR system for ETAPE French evaluation campaign : experiments on system combination using open-source recognizers. In *6th International Conference on Text, Speech and Dialogue (TSD'13)*.

CLEMENTS G. N. (1985). The geometry of phonological features. *Phonology*, **2**, pp. 225–252.

CONNOLLY J. (1997). Quantifying target-realization differences. Part I : Segments. *Clinical Linguistics & Phonetics*, **11**, pp. 267–287.

GAUVAIN J., LAMEL L., SCHWENK H., ADDA G., CHEN L. & LEFÈVRE F. (2003). Conversational telephone speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*.

GORYAINOVA M., GROUIN C., ROSSET S. & VASILESCU I. (2014). Morpho-Syntactic Study of Errors from Speech Recognition System. In *LREC'14*, p. 3050–3056.

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 *8th International Conference on Language Resources and Evaluation (LREC'12)*.

HEERINGA W. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. PhD dissertation, Rijksuniversiteit Groningen, Groningen.

HEERINGA W., NERBONNE J. & KLEIWEG P. (2002). Validating Dialect Comparison Methods. In *24th Annual Meeting of the Gesellschaft für Klassifikation (GFKL'02)*, p. 445–452 : Springer.

KONDRAK G. (2000). A New Algorithm for the Alignment of Phonetic Sequences. In *6th Applied Natural Language Processing Conference (ANLP'00)*, p. 288–295.

KONDRAK G. (2003). Phonetic Alignment and Similarity. *Computers and the Humanities*, **37 (3)**, pp. 273–291.

LEVENSHTEIN V. I. (1965). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Doklady Akademii Nauk SSSR*, **163**, pp. 845–848.

LUZZATI D., GROUIN C., VASILESCU I., ADDA-DECKER M., BILINSKI E., CAMELIN N., KAHN J., LAILLER C., LAMEL L. & ROSSET S. (2014). Human annotation of ASR error regions : Is "gravity" a sharable concept for human annotators ? In *9þth International Conference on Language Resources and Evaluation (LREC'14)*, p. 3050–3056.

MESTER R. A. & ITO J. (1989). Feature Predictability and Underspecification : Palatal Prosody in Japanese Mimetics. *Language*, **65 (2)**, pp. 258–293.

PARADIS C. & PRUNET J.-F. (1991). Introduction : Asymmetry and Visibility in Consonant Articulations. In C. PARADIS & J.-F. PRUNET, Eds., *The Special Status of Coronals : Internal and External Evidence*, p. 1–28. San Diego : Academic Press.

SAGEY E. C. (1986). *The Representation of Features and Relations in Non-linear Phonology*. PhD thesis, Massachusetts Institute of Technology.

SANTIAGO F., DUTREY C. & ADDA-DECKER M. (2015). Towards a Typology of ASR Errors via Syntax-Prosody Mapping. In *Errors by Humans and Machines in multimedia, multimodal and multilingual data processing (ERRARE'15)*.

VINTSYUK T. (1968). Speech Discrimination by Dynamic Programming. *Kibernetika*, **4**, pp. 81–88.

WALKER R. (1993). A Vowel Feature Hierarchy of Contrastive Specification. *Toronto Working Papers in Linguistics*, **12 (2)**, pp. 179–198.

Allophonie et position dans la syllabe : Indices acoustiques pour les consonnes latérales

Anisia Popescu[1] , Ioana Chitoran[1]
(1) Université Paris Diderot, Clillac-ARP, 5 Rue Thomas Mann, 75013 Paris, France
anisia.popescu@univ-paris-diderot.fr, ioana.chitora@univ-paris-diderot.fr

RESUME

L'article traite de la manifestation acoustique des consonnes latérales en anglais américain et en roumain en fonction de la position syllabique et de la complexité phonotactique. Nous avons considéré quatre types de mesures: valeurs formantiques, équations locus, ratio d'intensité et présence/absence de relâchements. Notre but est, d'une part, de classifier les allophones des deux langues considérées et d'autre part de déterminer les indices acoustiques des gestes articulatoires des consonnes latérales. Les résultats indiquent des différences importantes entre les deux langues. On montre que la distribution des allophones n'est pas binaire, mais graduée et que le statut du geste dorsal peut être considéré comme un marqueur de « degré de clarté ». On montre aussi que l'allophonie dépend de la position syllabique mais pas forcément de la complexité syllabique.

ABSTRACT

Acoustics of syllable position allophony: The case of lateral consonants. This study investigates the acoustic manifestation of syllable position allophony in lateral consonants in two languages – American English and Romanian. We compare several parameters: formant values, locus equations, intensity ratios and presence/absence of release bursts, in order to classify the allophones in the two languages studied, and to extract information about articulatory gestures from the acoustic data. The results reveal important cross-language differences. We show that the allophones in the two languages considered do not have a binary distribution, but rather a gradual one, and that the behavior of the dorsal gesture of the lateral could be considered a marker of "degree of clarity". We also show that the allophony is dependent on syllable position but not necessarily on syllable complexity.

MOTS-CLES : acoustique, allophonie, syllabe, latérales, coarticulation
KEYWORDS: acoustics, allophony, syllable, laterals, coarticulation

Introduction

La production des segments comportant deux gestes articulatoires, tels que les consonnes liquides, les nasales ou les glides, a été amplement étudiée. La variation allophonique des liquides a été étudiée à la fois d'un point de vue acoustique (Ladefoged et Maddieson 1995, Recasens 2012) et articulatoire (Proctor 2009, Proctor et Walker 2012 ; Marin et Pouplier 2014). Dans cet article on présente une étude acoustique des consonnes latérales dans deux langues (anglais-américain et

roumain) en considérant quatre paramètres : la structure formantique, le degré de coarticulation et deux nouveaux paramètres: l'intensité et la présence ou l'absence de relâchements.

Le but de l'article est d'approfondir nos connaissances de l'allophonie latérale en répondant à la question suivante : quelle est la manifestation acoustique du statut consonantique vs. vocalique du double geste en fonction de la position syllabique (attaque ou coda) et de la complexité phonotactique (attaque/coda simple ou complexe).

Composition gestuelle et allophonie des consonnes latérales

La composition gestuelle de la consonne latérale est bien documentée d'un point de vue articulatoire (Sproat et Fujimura 1993, Proctor 2009, Proctor et Walker 2012 ; Marin et Pouplier 2010,2014) : il est établi que /l/ implique deux gestes articulatoires distincts : celui d'une consonne apicale, représenté par un rehaussement de l'apex, et celui d'une voyelle postérieure, qui consiste d'une rétraction du dos de la langue. Nous avons étudié ici deux langues qui diffèrent par rapport à la présence/absence d'allophonie en fonction de la position de la syllabe : l'anglais américain (AA) et le roumain (RO).

En AA /l/ a deux allophones en fonction de la position dans la syllabe : /l/ clair en attaque et /l/ sombre en position coda. En RO /l/ n'a pas d'allophones, et se comporte comme /l/ clair en attaque et en coda (Marin et Pouplier 2010, Recasens 2012). La différenciation des deux implique des coordinations temporelles différentes des deux gestes articulatoires composants (Sproat et Fujimura 1993).

On connait moins bien, pourtant la manifestation acoustique de ce double statut consonantique/vocalique des deux gestes impliqués dans la production des consonnes latérales. Ainsi on propose de répondre à la question posée dans l'introduction en effectuant une analyse acoustique détaillée du /l/ dans les deux langues considérées.

Choix des paramètres acoustiques

Les paramètres acoustiques considérés dans l'analyse des consonnes latérales se rapportent principalement à leur structure formantique. On reprend tout d'abord la mesure utilisée par Sproat et Fujimura (1993) pour évaluer le degré de clarté (clair/sombre) de /l/ dans chaque langue. La différence des valeurs des deux premiers formants (F2-F1) peut déterminer le degré de clarté. Plus la différence F2-F1 est grande plus la consonne latérale est claire. En effet /l/ clair se caractérise par un F2 élevé (>1500Hz) et par un F1 bas (<400Hz) (Recasens 2012) alors que pour la variante sombre F2 baisse et F1 monte. A cette mesure on propose d'ajouter deux nouveaux paramètres : le ratio d'intensité et la présence ou l'absence de pics d'énergie durant /l/. Le ratio d'intensité est donné par l'intensité moyenne de /l/ divisée par l'intensité moyenne de la voyelle de la même syllabe. Etant donné que les deux segments (/l/ et /V/) sont adjacents, le temps est trop court pour que la valeur d'intensité soit influencée par d'autres facteurs. Souvent les /l/s présentent des pics d'énergie sous forme de relâchements. La Figure 1 illustre la présence de pics en attaque dans le mot anglais *lap* 'tour' et en coda dans le mot roumain *cal* 'cheval' :

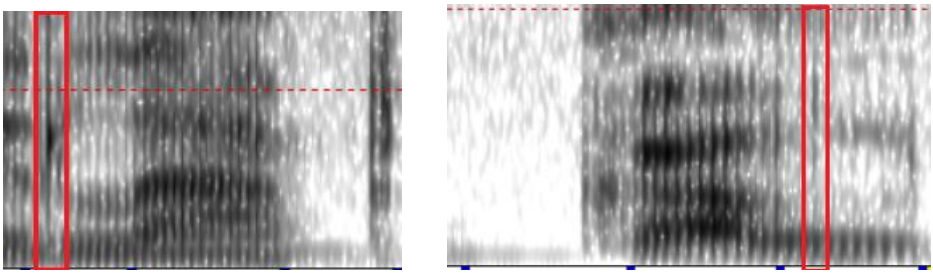


Figure 1: Spectrogramme lap *(AA) à gauche et* cal *(RO) à droite*

Le quatrième paramètre envisagé est l'équation locus, indice du degré de coarticulation entre la consonne latérale et la voyelle adjacente. Cette mesure pourra nous informer sur les relations acoustiques-articulatoires de la production des consonnes latérales. Les équations locus ont été introduites par Lindblom (1963) pour mesurer le degré de coarticulation et ont été reprises ultérieurement pour caractériser la coarticulation consonne-voyelle (C-V) (Krull 1989 ; Tabain 2000 ; Fowler et Brancazio 2000) et le lieu d'occlusion (Sussman 1993 ; Iskarous, Fowler et Whalen 2010). Les équations locus sont une régression linéaire faite à partir des valeurs du deuxième formant (F2) en fixant la consonne (/l/) et en variant le contexte vocalique (RO : /a i o u ɨ/ ; AA : /æ ɪ i ʌ ʊ/). Les mesures de F2 sont prises à l'état stable de la voyelle, c'est-à-dire au point de dérivée zéro, et au début de la transition formantique, correspondant à la limite C-V resp.V-C (Lindblom 1963). Dans le cas d'absence d'un état stable durant la voyelle nous avons pris la mesure au point du milieu du segment vocalique. Les mesures qui paramétrisent la régression linéaire sont la pente et l'intersection de la ligne de régression avec l'ordonnée (Krull 1989). Un troisième paramètre R^2 est utilisé correspondant au coefficient de régression. La pente a été utilisée comme paramètre acoustique pour mesurer le degré de coarticulation. Une pente zéro indique une coarticulation minimale. Une pente égale à l'unité indique une coarticulation maximale (Krull 1989). Plus la valeur F2 à la limite C-V (resp. V-C) est variable, plus le degré de coarticulation est fort.

On résume ici les quatre principaux paramètres:

1. F2 - F1 comme mesure de clarté : des valeurs plus basses indiquent un /l/ plus sombre

2. Ratio d'intensité - Moyenne d'intensité de /l/ sur moyenne d'intensité de la voyelle appartenant à la même syllabe : Un ratio plus proche de 1 correspond à une production plus vocalique de la liquide.

3. Présence vs. absence de pics d'énergie : On considère que la présence de relâchements indique un degré de constriction plus élevé, qui correspond à une production plus consonantique de la liquide.

4. Degré de coarticulation donné par les équations locus : une pente proche de 0 indique une coarticulation minimale, une pente proche de 1 indique une coarticulation maximale.

Hypothèses basées sur les mesures acoustiques

Etant données les mesures introduites dans la section précédente on émet les hypothèses suivantes pour les deux langues comparées.

1. F2-F1 :
   - AA : les valeurs F2-F1 seront plus élevées en position attaque qu'en coda, indiquant l'allophonie.
   - RO : les valeurs F2-F1 vont être similaires en attaque et coda indiquant l'absence d'allophonie.
2. Ratio d'intensité :
   - AA : une différence d'intensité est attendue entre attaque et coda avec une valeur se rapprochant de 1 en coda pour un /l/ sombre et des valeurs inférieures en attaque pour un /l/ clair.
   - RO : des valeurs similaires sont prévues en attaque et coda en absence d'allophonie.
3. Pics d'énergie (relâchements) :
   - AA : moins de pics seront présents en coda pour un /l/ plus vocalique et plus de pics en attaque pour un /l/ plus consonantique.
   - RO : pas de différence entre attaque et coda.
4. Equations locus : On s'attend à des degrés différents de coarticulation, par ordre croissant : le plus faible degré de coarticulation en AA coda, suivi par AA attaque, le plus fort degré de coarticulation en RO attaque et coda.

Les hypothèses sur les équations locus (4) sont basées sur les résultats articulatoires de Bladon & Al-Bamerni 1976, Tabain 2000, Proctor 2009, Proctor et Walker 2012 et Recasens 2012 : Si le geste dorsal de rétraction est une cible articulatoire de la consonne, alors cette cible sera atteinte et par conséquent la coarticulation avec la voyelle adjacente sera faible. Les deux allophones de /l/ (clair et sombre) impliquent tous les deux un geste dorsal, différant seulement par la coordination temporelle avec le geste apical. D'après Recasens 2012 /l/ sombre a un geste dorsal plus prononcé que /l/ clair et sera donc plus résistant à la coarticulation. On s'attend donc à un faible degré de coarticulation entre voyelle et consonne latérale en général, surtout en AA en position coda (/l/ sombre). Vu les résultats F2-F1, le /l/ du RO est encore plus clair que celui du AA. Si on prend en considération ces résultats, il est possible qu'on retrouve cette différence dans les équations locus, c'est-à-dire que le degré de coarticulation sera plus fort en RO qu'en AA.

Stimuli et enregistrements

Notre corpus se compose de mots monosyllabiques comportant des /l/ seuls ou en clusters en position attaque et coda :

Langue	Attaque simple	Coda simple	Attaque complexe	Coda complexe
AA	lɪp	sɪɬ	blɪp	sɪɬk
RO	lin	ʧil	plin	film

Tableau 1 : Exemples du corpus anglais-américain et roumain

Les contraintes phonotactiques des deux langues ont imposé le choix du contexte vocalique (RO : /a i u o ɨ/ et AA : /i ɪ æ ʊ ʌ/) et du type de clusters (RO : attaque /#pl-/, /#kl-/, coda /-ld#/, /-lg#/, /-lk#/,

/-lm#/, /-lt#/, -/lts#/ et AA : attaque /#bl-/, /#gl-/, : coda /-lp#/, /-lk#/, /-ld#/, /-lt#/). L'inventaire résultant est de 90 mots cibles en RO et 96 en AA. Chaque mot cible a été inséré dans une phrase porteuse et répété 5 fois dans des blocs randomisés. On a fait varier les phrases porteuses, tout en gardant le même nombre de syllabes totales (5), pour éviter des effets de cadence rythmique.
Au contraire des études précédentes, nous avons examiné des mots monosyllabiques avec /l/ en attaque et coda, en évitant les frontières morphologiques.
Huit locuteurs par langue ont été enregistrés avec un enregistreur Edirol 24 bit 96kHz Wave/MP3. Les locuteurs américains ont été enregistrés à Paris dans la cabine insonorisée de Paris Diderot. Ce sont des étudiants qui ont passé un trimestre à Paris dans un programme d'études universitaires. Les locuteurs roumains ont été enregistrés à Bucarest en Roumanie dans la salle insonorisée de l'Académie Roumaine. Les résultats présentés dans cet article sont des résultats préliminaires de toutes les mesures acoustiques portant sur les enregistrements de six locuteurs masculins, trois par langue. Les résultats ont été analysés statistiquement en utilisant des Effets Linéaires Mixtes avec '*langue'*, *'position dans la syllabe'* et *'complexité syllabique'* comme facteurs fixes et '*locuteur'* et '*mot'* comme facteurs aléatoires. Etant donnée la variation du contexte vocalique des stimuli des deux langues, nous avons rajouté l'*'aperture'* de la voyelle comme facteur aléatoire pour les mesures d'intensité. Ainsi tout effet crée par la différence d'intensité entre voyelles fermées et ouvertes n'est pas pris en compte. Pour les mesures des équations locus nous avons utilisé la fonction régression linéaire en R.

Méthodologie

Les mesures pour le degré de clarté et les équations locus consistant de valeurs formantiques ont été prises manuellement à partir des spectrogrammes au point milieu (état stable) du segment vocalique et de la liquide, ainsi qu'à la frontière des deux segments. Le critère de segmentation est la différence d'intensité au niveau du deuxième et troisième formant entre la voyelle et la consonne latérale. Un script Praat a été utilisé pour extraire les valeurs moyennes du segment /l/ et de la voyelle adjacente. Pour la détection des relâchements on a identifié et compté les /l/s comportant un ou plusieurs pics de relâchement, observés sur les oscillogrammes et les spectrogrammes, et on a ensuite calculé les pourcentages par rapport au nombre total de mots cibles. Le critère d'identification des pics est toute concentration verticale d'énergie qui crée une non-uniformité dans l'ensemble du segment (voir Figure 1).

Résultats et interprétations

Les résultats concernant le degré de clarté donné par les valeurs de F2-F1 (Figure 2) indiquent comme prévu un effet de *position dans la syllabe* pour l'AA ($p \sim 4.87e\text{-}4$). Cet effet confirme la présence d'allophonie dépendante de la position syllabique en AA. En RO l'absence d'effet confirme une absence d'allophonie. Un effet de *langue* ($p \sim 6.55e\text{-}4$) vient s'ajouter à celui de position syllabique. Le RO a des valeurs plus élevées de F2-F1, ce qui implique que le /l/ clair du RO reste considérablement plus clair que le /l/ clair de l'AA. Les différences inter-locuteurs sont plus prononcées en RO qu'en AA en général. Les valeurs de F2-F1 s'étendent sur un intervalle plus grand en RO qu'en AA. De plus en AA les valeurs F2-F1 en position attaque (/l/ clair) sont plus variables entre locuteurs que celles en coda (/l/ sombre). Ceci indique que plus la latérale est claire plus il y a des différences entre les locuteurs. Il est possible que la variabilité réduite du /l/ sombre soit liée au fait que pour ce type de /l/ chacun des deux gestes constitue une cible articulatoire. Le /l/ sombre est ainsi le moins variable et, on verra plus loin, le plus résistant a la coarticulation.

Pour les ratios d'intensité on trouve un effet de la langue (p~3.76e-3). Le RO a des ratios significativement plus petits. Cela signifie que le RO a une production plus consonantique du /l/ que l'AA. Ce résultat d'intensité soutient et renforce le résultat F2-F1 pour le degré de clarté. Toutefois on n'a pas trouvé d'effet de position syllabique dans aucune des deux langues par rapport à l'intensité. Alors qu'on ne s'attendait pas à trouver un effet dans le cas du RO, l'absence d'effet en AA est inattendue. Elle peut cependant s'expliquer par les pics d'énergie (Tableau 2) portant sur l'absence ou la présence de relâchements de /l/.

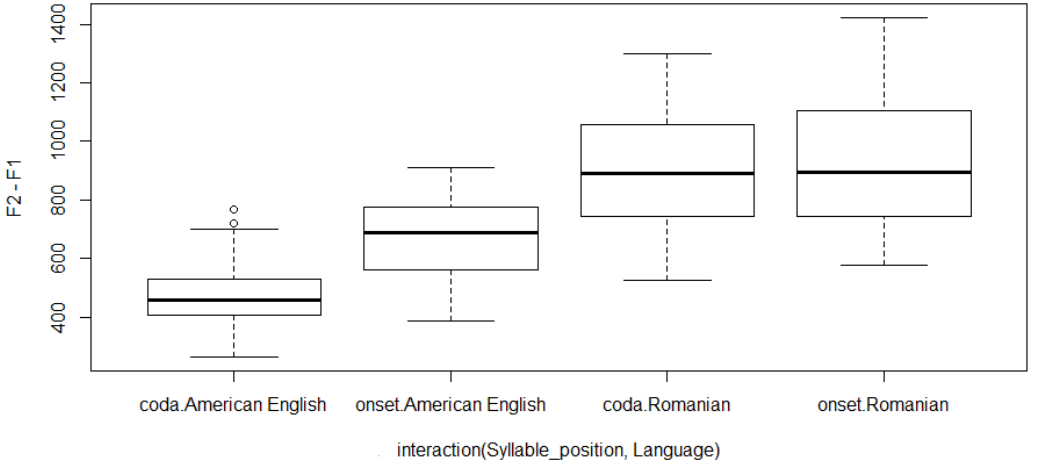


Figure 2: Valeurs F2-F1 variant par langue et position syllabique

Nous n'avons pas trouvé d'effet de *complexité syllabique* pour aucun des deux paramètres, ni F2-F1, ni ratio d'intensité. Les qualités intrinsèques de la consonne latérale dans les deux langues ne changent donc pas en fonction de la complexité syllabique.

Pour les pics d'énergie, en RO on a trouvé des pourcentages similaires de présence de relâchements en attaque et en coda, indiquant encore une fois l'absence d'allophonie. Pour l'AA, une grande différence entre attaque et coda confirme l'allophonie soutenue par les résultats F2-F1. On remarque ici que la distribution des pics est uniforme parmi les locuteurs.

% relâchements	/l/ ATTAQUE	/l/ CODA
AA	66%	12.5%
RO	44%	42%

Tableau 2 : % de relâchements de /l/

Le pourcentage important de pics en attaque pour l'AA peut être considéré comme un facteur perturbant des résultats d'intensité. Nous avons mesuré l'intensité moyenne sur toute la durée du segment /l/, donc la présence de relâchements (qui sont des pics d'énergie) aurait pu augmenter la valeur totale de l'intensité moyenne correspondant à la consonne latérale. L'augmentation de l'intensité moyenne de /l/ rapproche de 1 le rapport (intensité /l/ / intensité /V/) ce qui entraine un manque d'effet de position syllabique en AA. Dans le cas du RO, par contre cela ne change pas les résultats car la présence de pics de relâchements est comparable en attaque et en coda. Ainsi, si la

présence de relâchement augmente la valeur moyenne d'intensité, en RO l'augmentation sera la même pour les deux positions syllabiques. .

Même si le pourcentage élevé de relâchements en attaque en AA peut expliquer le manque d'effet de position syllabique pour l'intensité, il reste tout de même difficile à interpréter relatif à l'intensité. Dans nos hypothèses nous avons interprété la présence de relâchements comme l'indication d'une production plus consonantique de la latérale. Le RO ayant une production plus consonantique (comme établi par l'effet de langue sur les valeurs d'intensité), on s'attendait à un pourcentage de relâchements plus élevé en RO qu'en AA. Or ce n'est pas le cas. Ainsi il est possible que la présence de relâchements indique principalement une constriction plus forte, que ce soit une constriction apicale ou dorsale. Cela veut dire qu'il ne serait plus pertinent d'attribuer une qualité « consonantique » ou « vocalique » au /l/ en fonction de la présence de relâchements. En effet, on a pu trouver des pics de relâchement même dans des voyelles fermées comme /i/ et /ɨ/. Les résultats concernant la présence de relâchements ne donnent donc d'information fiable que sur l'allophonie : ils confirment l'allophonie en AA et son absence en RO.

Finalement, le Tableau 3 et la Figure 3 montrent les résultats des équations locus, l'indice de coarticulation entre la consonne liquide et le noyau vocalique tautosyllabique. On trouve que indépendamment de la position et de la complexité syllabique, en RO le /l/ a un fort degré de coarticulation et en AA un faible degré de coarticulation.

	Attaque simple		Attaque complexe		Coda simple		Coda complexe	
Langue	AA	RO	AA	RO	AA	RO	AA	RO
Pente	0.4609	0.705	0.3622	0.7238	0.5498	0.6325	0.4809	0.7238
Intersection	506.77	384.01	663.71	315.09	318.37	473.5	506.77	315.09
R^2	0.34	0.93	0.12	0.94	0.89	0.75	0.34	0.94

Tableau 3 : Paramètres équations locus : pentes, intersection et R^2 en fonction de la langue et de la position syllabique

Les résultats pour l'AA sont en partie confirmés. Le degré de coarticulation en AA est faible (pente plus éloignée de 1) par rapport au RO, mais il n'y a pas de différence entre attaque et coda. Ce résultat est attribué à la présence d'une cible dorsale plus importante du /l/ en AA en général, indépendamment de sa position syllabique.

Par contre, pour le RO on peut attribuer le degré de coarticulation relativement fort à une cible dorsale réduite par rapport à la cible apicale. On peut dire que les deux langues diffèrent par rapport au poids relatif des deux gestes. En effet si en RO le geste dorsal était une cible articulatoire du /l/, alors la variation du deuxième formant aurait dû être plus faible comme dans le cas de l'AA. Ceci confirme l'hypothèse d'une production plus consonantique du /l/ en RO qu'en AA.

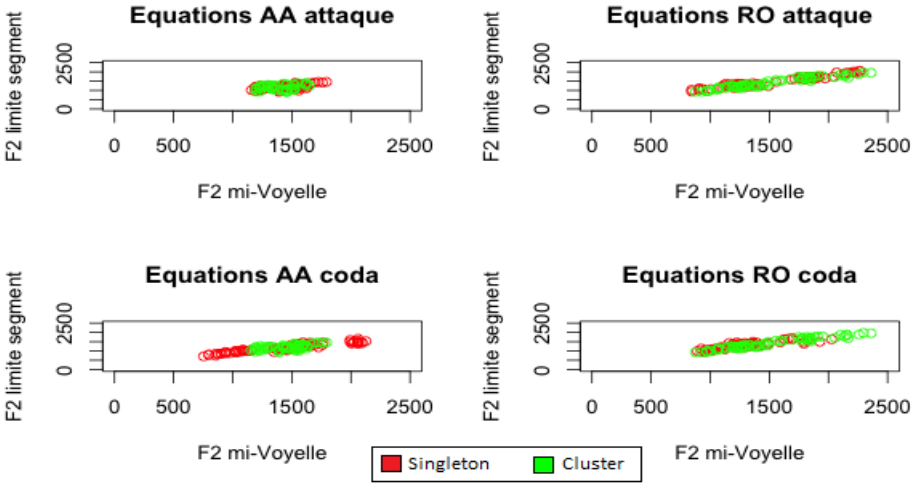


Figure 3 : Valeurs F2 – en début de transition (limite segment) *en fonction des valeurs* F2 – à mi voyelle *en attaque et coda pour AA (gauche) et RO (droite*)

Conclusion

L'article propose une analyse acoustique des réalisations des consonnes latérales en deux langues avec et sans allophonie syllabique, prenant en compte la double composition gestuelle en fonction de la position syllabique et de la complexité phonotactique. Nous avons analysé quatre paramètres : F2-F1, le ratio d'intensité, les pics d'énergie en tant que relâchement, les équations locus comme indice de coarticulation avec le noyau vocalique. Dans un premier temps, la comparaison de deux langues dans la même analyse nous a permis d'obtenir de nouvelles informations typologiques. Les mesures de degré de clarté (F2-F1) montrent que l'allophonie /l/ clair - /l/ sombre n'a pas une distribution binaire, mais plutôt graduelle. La variante claire de l'anglais a un degré intermédiaire de clarté lorsqu'on la compare à la variante claire du roumain. Nos résultats s'accordent à ceux de Recasens 2012. Nous proposons que le /l/ de l'anglais a une cible dorsale plus importante, qui se manifeste dans les résultats acoustiques, alors que le /l/ du roumain a un geste apical comme cible principale. Le /l/ sombre de l'AA est aussi le moins variable entre locuteurs, ce qui pourrait s'expliquer par le fait que sa production demande plus de précision, surtout pour le geste dorsal. Notre hypothèse s'accorde aussi avec des résultats préliminaires de données ultrasons (King, en cours de préparation) qui montrent que les locuteurs d'AE ont bien un composante dorsale dans la production des deux allophones de /l/ (sombre et clair) contrairement aux locuteurs d'anglais britannique qui ont un geste dorsal réduit. Ces résultats ont des implications importantes pour la modélisation inter-langue des consonnes latérales.

Finalement on a montré que la position syllabique est un facteur qui influence la présence d'allophonie alors que la complexité syllabique ne l'est pas, du moins pour les quatre paramètres considérés ici. Toutefois, on sait que, au-delà de l'allophonie, la complexité syllabique a une influence sur la coordination temporelle entre consonnes latérales et voyelles d'une même syllabe (Marin et Pouplier 2010, 2014). Des futurs travaux vont d'une part intégrer plus de locuteurs des deux langues et examiner des hypothèses plus précises portant sur la coordination temporelle du double geste des consonnes liquides.

Références

BLADON R.A.W., AL-BAMERNI A. (1976). Coarticulatory resistance in English /l/. *Journal of Phonetics,* 3, 137-150.

FOWLER C., BRANCAZIO L. (2000). Coarticulation resistance of American English consonants and its effects of transconsonantal vowel-to-vowel coarticulation. *Lang Speech,* 43, 1-41.

ISKAROUS K.,FOWLER C.,WHALEN D.H. (2010).Locus equations are an acoustic expression of articulator synergy. *J. Acoust. Soc. Am,* 127, 2021-2032

KING H. (En cours de préparation).The dark side of the tongue: étude de faisabilité de l'utilisation de l'imagerie à ultrasons pour l'acquisition du /l/ sombre chez les apprenants francophones de l'anglais. Mémoire de Master 2, Université Paris 7-Paris Diderot.

KRULL D. (1989).Second formant locus patterns as a measure of consonant-vowel coarticulation. *PERILUS,* 5, 43-61.

LADEFOGED D., MADDIESON I. (1995). The sounds of the World's Languages. Blackwell Publishing

LINDBLOM B. (1963).Spectographic study of vowel reduction. *J. Acoust. Soc. Am.,* 121, 1773-1781.

MARIN S.,POUPLIER M. (2010). Temporal Organisation of Complex Onsets and Codas In American English: Testing the Predictions of a Gestural Coupling Model. *Motor Control* 14, 380-407.

MARIN S., POUPLIER M. (2010). Articulatory synergies in the temporal organization of liquid clusters in Romanian. *Journal of Phonetics,* 24, 24-36.

PROCTOR M., (2009). *Gestural Characterization of a Phonological Class: The Liquids*. Thèse de Doctorat : Yale University.

PROCTOR M., WALKER R. (2012). Articulatory bases of sonority in English liquids. *The Sonority Controversy,* 18, 289.

RECASENS D. (2012). A cross-language acoustic study of initial and final allophones of /l/. *Speech Communication,* 54, 368-383.

SPROAT R., FUJIMURA O. (1993). Allophonic variation in American English /l/ and its implications for phonetic interpretation. *Journal of Phonetics,* 21, 291-311.

SUSSMAN H., HOEMEKE K., AHMET F. S. (1993). A cross-linguistic investigation of locus equations as a phonetic descriptor for place articulation. *J. Acoust. Soc. Am,* 94, 1256-1268.

TABAIN M. (2000). Coarticulation in CV syllables: A comparison of locus equation and EPG data. *Journal of Phonetics,* 28, 137-159.

Analyses acoustiques des monophtongues du luxembourgeois produites dans la parole lue

Tina Thill
Institut de langue et de littératures luxembourgeoises, Université du Luxembourg
11, porte des Sciences, L-4366 Esch-sur-Alzette, Luxembourg
Laboratoire de Phonétique et Phonologie, Université Sorbonne Nouvelle – Paris 3
19, rue des Bernardins, F-75005 Paris, France
tina.thill@ymail.com, tina.thill@uni.lu

RESUME

Cet article présente une analyse acoustique de 12 monophtongues du luxembourgeois produites par des locuteurs de la région centrale du Grand-Duché de Luxembourg. Cette analyse fait partie du travail empirique de notre thèse de doctorat sur les productions natives et non natives des voyelles du luxembourgeois. A partir des données de 10 locuteurs natifs, nous analysons les valeurs de la durée et des trois premiers formants des paires de voyelles longues et brèves opposées [iː]-[i], [eː]-[e], [aː]-[ɑ], [oː]-[ɔ], [uː]-[u] et de l'allophone [ɛː] réalisée lorsqu'elle est suivie d'un /r/. Les analyses montrent que (i) les voyelles longues et brèves se distinguent tant par la durée acoustique que par le timbre, (ii) la voyelle semi-ouverte [ɛː] suivie d'un /r/ vocalisé tend à se diphtonguer.

ABSTRACT

Acoustic analyses of Luxembourgish monophthongs produced in reading speech

In this article, we investigate the production of 12 monophthongs produced by 10 native speakers of Luxembourgish of the central region of the Grand-Duchy of Luxembourg. The analyses are part of the empirical work of our PhD thesis, which investigates native and non-native productions of Luxembourgish vowels. The aim of this article is to analyze formant and duration values of the pairs of contrastive long and short vowels [iː]-[i], [eː]-[e], [aː]-[ɑ], [oː]-[ɔ], [uː]-[u] and the mid-open [ɛː], which is only realized when followed by /r/. The analyses show that (i) both duration and vowel quality are important for the distinction between long and short vowels, (ii) [ɛː] followed by /r/ is characterized by diphthongization.

MOTS-CLÉS : luxembourgeois, phonétique acoustique, monophtongues, parole lue
KEYWORDS: Luxembourgish, acoustic phonetics, monophthongs, reading speech

1 Introduction

A l'origine un dialecte francique mosellan, le luxembourgeois (*Lëtzebuergesch*), parlé par environ 400.000 locuteurs (FEHLEN, 2009), est une langue encore peu étudiée sur le plan phonétique. A côté du contact important entre le luxembourgeois, l'allemand et le français au Grand-duché de

Luxembourg, il existe de nombreuses variétés régionales. Actuellement, le luxembourgeois connaît aussi un essor dans l'enseignement en tant que langue étrangère (LLE), ce qui pousse à vouloir décrire la langue pour des besoins pédagogiques afin de fournir un matériel adéquat pour son apprentissage (TROUVAIN ET GILLES, 2009). L'enseignement du LLE se base sur la variété parlée la plus répandue, représentée par le parler de la région centrale du Luxembourg (GILLES, 1999). L'objectif de cet article est d'investir les caractéristiques acoustiques des voyelles de cette région afin d'approfondir les descriptions phonétiques et d'obtenir des valeurs de référence pour les formants qui puissent servir pour des études comparatives. Cette approche a déjà été menée par GEORGETON ET AL. (2012) dans le cadre d'un projet sur le français langue étrangère.

Les travaux de KEISER-BESCH (1976) et GOUDAILLIER (1987) figurent parmi les premières études phonétiques et phonologiques sur le luxembourgeois. Des études plus récentes ont remis en question l'ancien schéma phonologique et ont fourni de nouvelles descriptions sur les voyelles produites dans la parole lue (GILLES ET TROUVAIN, 2013 ; NISHIDE, 2014). D'autres études se sont penchées sur les réalisations des voyelles dans la parole journalistique pour la reconnaissance automatique de la parole (ADDA-DECKER, 2014). D'une manière générale, ces travaux appellent à un besoin de recherche plus approfondie sur le luxembourgeois.

Le travail présenté dans cet article part sur le schéma vocalique proposé par GILLES ET TROUVAIN (2013) et basé sur les données d'un locuteur de la région centrale du Luxembourg. Selon ce système, le luxembourgeois possède 14 monophtongues, 8 diphtongues et 9 emprunts du français et de l'allemand (cf. **TABLEAU 1**). Parmi les monophtongues, nous constatons qu'il existe des paires de voyelles longues et brèves [iː]-[i], [eː]-[e], [aː]-[ɑ], [oː]-[ɔ], [uː]-[u] qui permettent de distinguer entre les paires minimales suivantes :

(1)	*wiiss*	*[viːs]*	*« grandis »*	–	*Wiss*	*[vis]*	*« pré »*
(2)	*Scheek*	*[ʃeːk]*	*« étui »*	–	*schéck*	*[ʃek]*	*« envoie ! »*
(3)	*Pap*	*[paːp]*	*« colle »*	–	*Papp*	*[pap]*	*« père »*
(4)	*Mooss*	*[moːs]*	*« mesure »*	–	*Moss*	*[mɔs]*	*« nana »*
(5)	*Muuss*	*[muːs]*	*« minou »*	–	*muss*	*[mus]*	*« dois »*

Exemples d'opposition entre les paires de voyelles longues et brèves par des paires minimales

Ces paires minimales constituent un point de départ pour notre analyse. En effet, une analyse sur le degré de distinction acoustique entre ces paires permettrait de d'approfondir les descriptions sur les caractéristiques de ces voyelles dans la parole. Notons que dans la liste des monophtongues, nous retrouvons également deux voyelles semi-ouvertes [ɛː] et [æ] ainsi que deux voyelles centrales [ə] et [ɐ]. La voyelle [ɛː] considérée comme un allophone de /eː/ qui se réalise seulement lorsqu'il est suivi d'un /r/, souvent vocalisé chez les jeunes locuteurs (ex. *Päerd* /peːrt/ « cheval »). Ce cas de variation, que l'on retrouve aussi en allemand (KOHLER, 1977), n'est cependant pas plus explicité dans la littérature et exige également une observation analytique particulière et une confirmation appuyée par des résultats. Pour enrichir l'analyse, nous incluons également les valeurs de [æ].

Dans le but d'approfondir et d'étendre les descriptions de GILLES ET TROUVAIN (2013) avec plus de locuteurs, le sujet de cet article se concentre sur deux questions : (1) les paires de voyelles longues et brèves se distinguent-elles acoustiquement par la durée et le timbre ? (2) Quel effet le /r/ produit-il sur la voyelle antérieure semi-ouverte [ɛː] avant /r/ (ex. *Päerd* [pɛːrt]) ? En nous basant sur les dits de la littérature, nous prononçons deux hypothèses : (i) les voyelles phonologiquement longues et brèves se distinguent par la durée acoustique et par le timbre ; (ii) la vocalisation du /r/ engendre une

diphtongaison du [ɛː] (ex. *Päerd* [pɛːət]). Nous nous attendons donc, d'une part, à une opposition importante entre les voyelles longues et brèves au niveau de la durée mais aussi au niveau du timbre et, d'autre part, à un changement de qualité vocalique pour [ɛː].

LE SYSTEME VOCALIQUE DU LUXEMBOURGEOIS		
Les voyelles natives		Les voyelles empruntées
Les monophtongues	Les diphtongues	
[iː], [i], [eː], [e], [ɛː], [æ], [aː], [ɑ], [oː], [ɔ], [uː], [u], [ə], [ɐ]	[iə], [əi], [uə], [əu], [æi], [æu], [ɑi], [ɑu]	[yː], [y], [øː], [œː], [œ], [ɛ̃], [ɑ̃], [ɔ̃], [ɒɪ]

TABLEAU 1: Les voyelles du luxembourgeois, d'après le système de GILLES ET TROUVAIN (2013).

Tout d'abord, nous présentons la méthodologie et le corpus, avant d'enchaîner sur les analyses et la discussion des résultats.

2 Méthode et données

En luxembourgeois, les voyelles peuvent se trouver en position tonique ou atone (GILLES ET TROUVAIN, 2013). De ce fait, et en raison d'un manque de description prosodique sur l'accent en luxembourgeois, nous analysons les monophtongues toniques dans un environnement consonantique. Les données proviennent du corpus Lëtz-Co, que nous avons développé dans le cadre de notre thèse de doctorat qui investit la comparaison entre les productions des voyelles par des locuteurs natifs et des apprenants français menée au sein de l'Institut de langue et de littératures luxembourgeoises à l'Université du Luxembourg et au Laboratoire de Phonétique et Phonologie à l'Université Sorbonne Nouvelle – Paris 3. Les enregistrements de mots lexicaux monosyllabiques et dissyllabiques en contexte isolé (ex. *Taass* « tasse ») ont permis d'obtenir les données pour les voyelles longues et brèves. Celles pour la voyelle semi-ouverte ont été obtenues à travers des enregistrements de cinq phrases à trous, où les participants étaient censés compléter le mot manquant de la phrase à l'aide d'une image (ex. *Nom Ausrett setze mir eist [Päerd] op d'Wiss.* « Après la promenade nous mettons notre cheval au pré. »). Dans l'exemple donné, le mot entre crochets représente le mot à trouver à l'aide d'une image. Cette procédure a permis d'enregistrer la séquence [ɛː] avant /r/ en évitant l'influence de l'orthographe et en favorisant une production spontanée de cette séquence. Le luxembourgeois n'étant pas enseigné à l'école, les locuteurs natifs n'ont pas l'habitude de lire selon les conventions orthographiques établies pour la langue.

Nous avons enregistré cinq hommes et cinq femmes âgées entre 24 et 59 ans (âge noté au moment des enregistrements en 2014). Les participants provenaient de la région centrale du Grand-Duché de Luxembourg et ils n'avaient pas vécu dans une autre région du Luxembourg. Ils ont été enregistrés dans une pièce calme avec un dictaphone Sony PCM-D50 et un micro casque Sennheiser HSP4 en mono et réglé à 44.1 kHz de fréquence d'échantillonnage. Chaque séance a duré environ 30 minutes.

Les voyelles ont été segmentées et annotées manuellement et les valeurs acoustiques (durée, F1, F2, F3) extraites avec un script sur le logiciel PRAAT (BOERSMA ET WEENINK, 2013). Les mesures ont été prises au milieu de la voyelle pour éviter l'influence du contexte phonémique. Les données ont été vérifiées et corrigées manuellement en cas d'erreur de détection des formants. En tout, nous disposons de 499 voyelles longues et brèves, dont 30 [iː], 70 [i], 40 [eː], 20 [e], 70 [aː], 70 [ɑ], 59 [oː], 70 [ɔ], 30 [uː], 40 [u], et 113 voyelles semi-ouvertes, dont 50 [ɛː] et 63 [æ]. En total, nous analysons 612 voyelles.

3 Analyses

Les analyses portent sur les valeurs de F1, F2, F3 et de la durée. Dans le but d'observer les réalisations vocaliques en détail, nous analysons les productions de chaque locuteur.

3.1 Moyennes et triangles vocaliques

Dans la littérature, les voyelles du luxembourgeois sont décrites à travers les deux premiers formants (GILLES ET TROUVAIN, 2013 ; NISHIDE, 2014), F1 étant associé au degré d'aperture et F2 au mouvement de la langue (DELATTRE *ET AL.*, 1952). Nous proposons d'inclure F3, associé à la labialité, et qui donne des informations supplémentaires sur les caractéristiques des voyelles (VAISSIERE, 2011). Dans cette section, nous présentons les valeurs de référence sous forme de moyennes pour les trois premiers formants (valeurs en Hertz) des voyelles longues et brèves et des voyelles semi-ouvertes. Le **TABLEAU 2** illustre les moyennes et les écart-types des valeurs de F1, F2 et F3 extraites au milieu des segments.

Les différences entre les valeurs de F3 que nous pouvons observer dans le **TABLEAU 2** justifient la prise en compte du troisième formant dans notre analyse. Par exemple, [iː]-[i] se distinguent au niveau de F2, mais aussi au niveau de F3, où la différence est visible à travers un écart de 282 Hz chez les hommes et de 456 Hz chez les femmes. L'écart est plus important pour les voyelles antérieures fermées. En effet, [eː]-[e] se distinguent à travers 318 Hz chez les hommes et 506 Hz chez les femmes, tandis que [oː]-[ɔ] par exemple affichent une différence de 64 Hz chez les hommes et de 31 Hz chez les femmes.

	HOMMES			FEMMES		
VOYELLES	F1	F2	F3	F1	F2	F3
iː	250 (21)	2123 (130)	2981 (131)	275 (29)	2551 (84)	3547 (276)
i	304 (25)	1999 (128)	2699 (152)	361 (66)	2387 (153)	3091 (282)
eː	323 (23)	2128 (125)	2760 (120)	403 (51)	2494 (127)	3150 (202)
e	460 (26)	1767 (85)	2442 (155)	513 (47)	2085 (110)	2644 (199)
ɛː	540 (56)	1809 (153)	2505 (199)	631 (72)	2016 (197)	2727 (143)
æ	641 (55)	1557 (129)	2319 (220)	748 (83)	1724 (128)	2666 (156)
aː	732 (77)	1379 (113)	2363 (230)	876 (68)	1523 (145)	2710 (145)
ɑ	629 (53)	1066 (103)	2491 (199)	708 (53)	1150 (83)	2630 (273)
oː	395 (55)	676 (143)	2748 (363)	409 (59)	791 (86)	2702 (244)
ɔ	474 (82)	799 (135)	2684 (292)	534 (44)	958 (105)	2751 (265)
uː	295 (65)	762 (153)	2610 (330)	303 (36)	801 (68)	2491 (152)
u	342 (43)	1024 (202)	2497 (364)	379 (62)	1019 (193)	2556 (124)

TABLEAU 2 : Les moyennes et écart-types des valeurs (Hz) de F1, F2 et F3 de 603 monophtongues en position tonique produites par 10 locuteurs natifs (5 hommes et 5 femmes).

Le triangle vocalique rend compte de la place des voyelles dans l'espace acoustique en fonction de F1 et F2. Sur la **FIGURE 1**, nous constatons un rapprochement entre [iː], [i] et [eː] et une valeur de F1 plus élevée pour [e]. La réalisation de [ɛː] est variable : chez la locutrice (à droite), [ɛː] est environ équidistant par rapport à [e] et [æ], tandis que chez le locuteur (à gauche), la voyelle est plus ouverte. Quant aux voyelles ouvertes, nous constatons, comme attendu, que [æ] est plus proche de [aː] que [ɑ]. La voyelle [oː] se rapproche considérablement de [ɔ], à l'image aussi de [uː] et [u]. De même, comme attendu, les voyelles brèves [i], [e] et [u] sont plus centrales que leurs opposées longues.

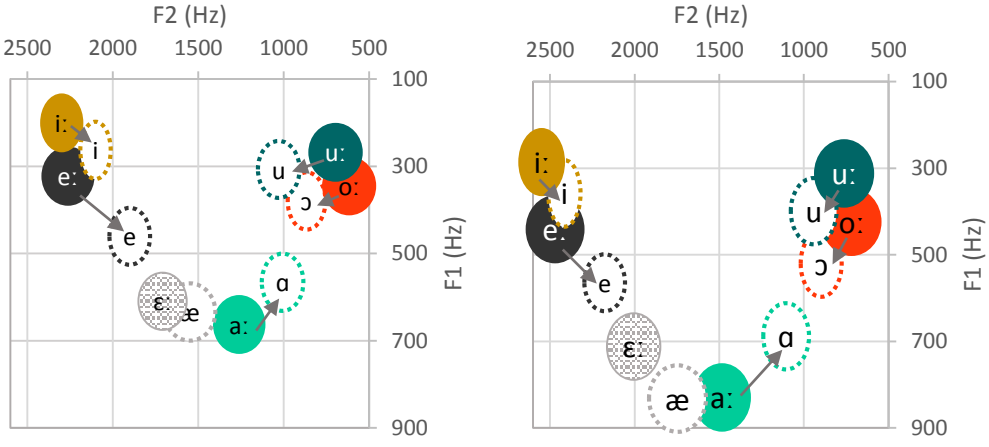


FIGURE 1 : Les triangles vocaliques d'un homme (à gauche) et d'une femme (à droite) représentant les monophtongues en fonction de F1 et F2 (valeurs non normalisées). Le lien entre les voyelles opposées par la longueur est indiqué par une flèche. Notons que la taille des ellipses n'est pas en rapport avec la dispersion des voyelles observées.

3.2 Voyelles longues et brèves

Afin de comparer les productions individuelles des 10 locuteurs, nous avons calculé les distances moyennes entre les paires de voyelles contrastives longues et brèves en fonction de F1, F2 et F3. Nous avons d'abord converti les valeurs de HERTZ en BARK pour réduire l'impact dû aux différences physiologiques entre les locuteurs dans les groupes des hommes et des femmes : cette procédure permet aussi de regrouper les voyelles acoustiquement proches sur une échelle auditive (GENDROT, 2013). Cette normalisation est effectuée à travers la formule suivante proposée par TRAUNMÜLLER (1990) :

$$Z_i = \frac{26{,}81}{\left(\frac{1+1960}{F_i}\right)} - 0{,}53$$

Dans cette formule, F_i indique la valeur de chaque formant soumis à la normalisation et Z_i correspond au résultat du calcul. Les distances ont été calculées avec la formule mathématique suivante :

$$\overline{D_{eucl}} = \sqrt{(Z_1(\mathrm{A}) - Z_1(\mathrm{B}))^2 + (Z_2(\mathrm{A}) - Z_2(\mathrm{B}))^2 + (Z_3(\mathrm{A}) - Z_3(\mathrm{B}))^2}$$

Le calcul de la distance euclidienne permet d'obtenir l'intervalle entre deux points sur un espace tridimensionnel. L'élément élevé au carré indique la soustraction de la valeur d'un formant Z de la voyelle A par celle de la voyelle B. Les distances ont été calculées à partir des moyennes en BARK pour les voyelles opposées [iː]-[i], [eː]-[e], [aː]-[ɑ], [oː]-[ɔ] et [uː]-[u].

Les voyelles longues et brèves sont bien distinguées au niveau du timbre, mais nous constatons que le degré de distinction varie en fonction des locuteurs. La **FIGURE 2** illustre globalement des écarts moins importants entre les paires de voyelles contrastives [iː]-[i] et [oː]-[ɔ] qu'entre les autres voyelles. La distinction entre [eː]-[e], [aː]-[ɑ] et [uː]-[u] est flagrante : l'écart entre [aː]-[ɑ] est particulièrement élevé chez PN3 et PN5 ; [uː]-[u] sont bien distinguées par tous les locuteurs, sauf par PN6, qui affiche un écart réduit entre ces deux voyelles. Chez les femmes, les distances entre les

voyelles se caractérisent par une plus grande variabilité. Ainsi, les distances chez PN6 tendent à être plus faibles, plus spécifiquement pour [iː]-[i] et [uː]-[u] : ces dernières sont par contre bien distinguées par les autres locutrices. Le degré de distinction est particulièrement élevé pour [aː]-[ɑ] chez PN3 et PN5, qui distingue également fortement les voyelles [oː]-[ɔ].

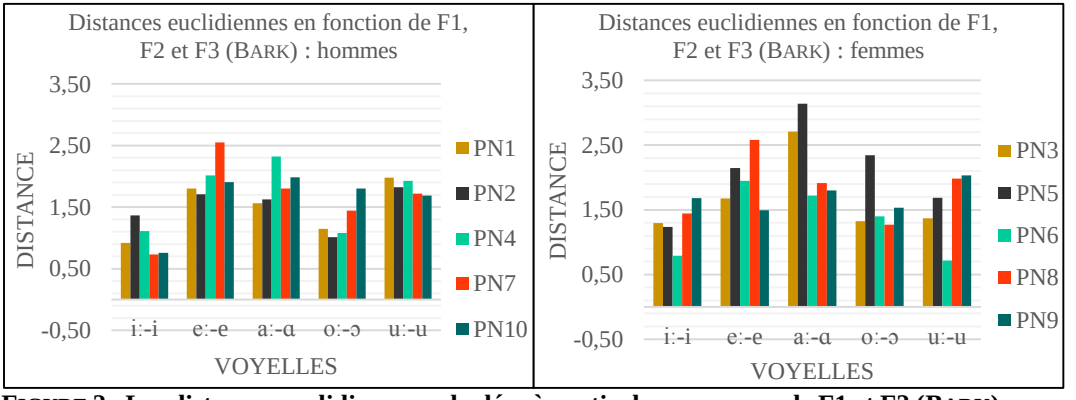


FIGURE 2 : Les distances euclidiennes calculées à partir des moyennes de F1 et F2 (BARK) pour les voyelles phonologiquement longues et brèves produites par les hommes et les femmes.

Les distances présentées sur ces graphiques suggèrent que certaines voyelles opposées se distinguent plus fortement que d'autres à travers les trois premiers formants. Pour détailler cet aspect, nous avons effectué des tests T indépendants, en incluant la durée. Nous avons normalisé les valeurs formantiques avec la méthode de LOBANOV pour réduire les différences dues aux facteurs sociologiques entre les locuteurs (THOMAS ET KENDALL, 2007) et pour pouvoir ainsi effectuer les tests sur les moyennes de l'ensemble des locuteurs. Les résultats des tests montrent une différence significative au niveau des trois premiers formants et de la durée entre les paires de voyelles contrastives [iː]-[i] (F1 : t = 7,65, p < 0,05 ; F2 : t = -4,98, p < 0,05 ; F3 : t = -7,12, p < 0,05 ; Durée : t = -14,96, p < 0,05), [eː]-[e] (F1 : t = 11,57, p < 0,05 ; F2 : t = -11,94, p < 0,05 ; F3 : t = -8,46, p < 0,05 ; Durée : t = -14,22, p < 0,05) et [aː]-[ɑ] (F1 : t = -11,33, p < 0,05 ; F2 : t = -17,81, p < 0,05 ; F3 : t = -0,77, p = 0,44 ; Durée : t = -12,77, p < 0,05). La différence n'est pas significative au niveau de F3 pour [oː]-[ɔ] (F1 : t = 7,60, p < 0,05 ; F2 : t = 6,34, p < 0,05 ; F3 : t = -0,51, p = 0,614 ; Durée : t = -11,41, p < 0,05) et [uː]-[u] (F1 : t = 7,60, p < 0,05 ; F2 : t = 6,34, p < 0,05 ; F3 : t = -0,51, p = 0,614 ; Durée : t = -11,41, p < 0,05). Ces résultats montrent l'importance du timbre et de la durée pour la production des paires de voyelles contrastives longues et brèves.

3.3 Réalisation de [ɛː] devant /r/

D'après la littérature, nous savons qu'un contexte uvulaire abaisse F2, notamment en finale de mot (GENDROT, 2013). Or, il est établi que le /r/ après [ɛː] à tendance à se vocaliser en luxembourgeois (GILLES ET TROUVAIN, 2013). Nous observons d'abord si l'abaissement de F2 de la voyelle a aussi lieu avec la vocalisation du /r/ et ensuite, s'il engendre une diphtongaison de la voyelle.

Pour observer la vocalisation du /r/, nous avons extrait les valeurs acoustiques de [ɛː] à sept points des segments. En moyenne, la voyelle a une durée de 120 ms. Les valeurs formantiques indiquent un mouvement dynamique de [ɛː], qui se traduit par un abaissement de F2 et une postériorisation de F1. (cf. **FIGURE 3**). Chez les hommes, la voyelle démarre à 1904 Hz au niveau de F2, puis engage abaissement progressif qui aboutit à 1485 Hz. Chez les femmes, F2 démarre à 2170 Hz et s'abaisse durant la production de la voyelle jusqu'à une centralisation à 1613 Hz ; F1 monte progressivement, de sorte qu'à la sixième mesure, l'espace entre les formants est pratiquement équidistant. Le

mouvement formantique de [ɛː] suggère un aspect dynamique de la voyelle (cf. **FIGURE 4**). Au début de sa production, la voyelle démarre au niveau de [e], puis descend vers [æ] et remonte vers le centre du triangle. Le mouvement dynamique de [ɛː] résulte en une diphtongaison. Il se peut cependant que ce résultat soit en partie influencé par les transitions formantiques au début et à la fin de la voyelle.

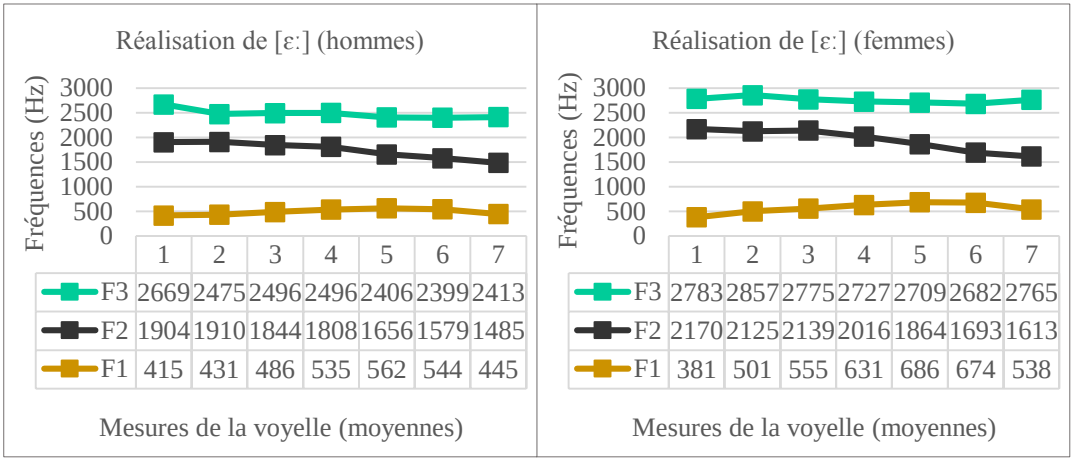


FIGURE 3 : La voyelle [ɛː] en fonction de sept mesures de F1, F2 et F3 (moyennes des valeurs de 25 voyelles par groupe).

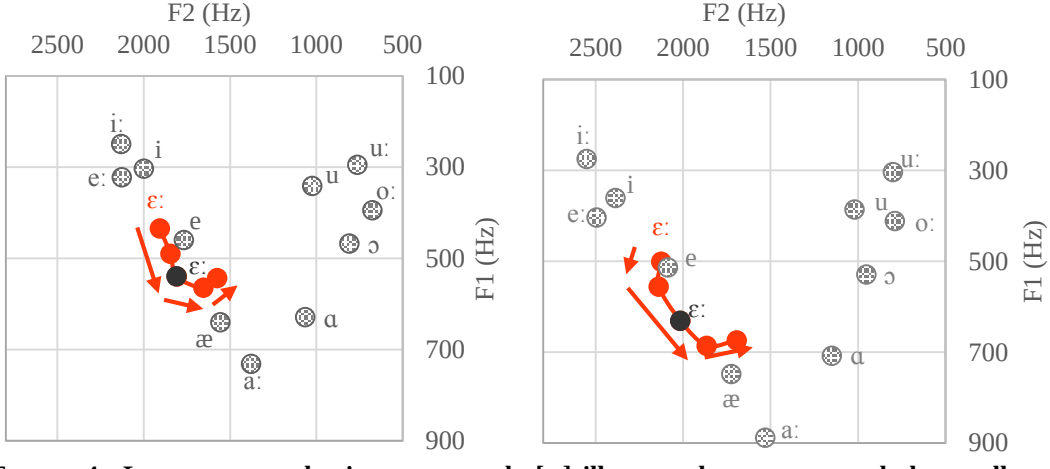


FIGURE 4 : Les moyennes de cinq mesures de [ɛː] illustrant le mouvement de la voyelle en fonction de F1 et F2 chez les hommes (à gauche) et les femmes (à droite).

4 Discussion et perspectives

Dans cet article, nous avons investi les caractéristiques acoustiques de 12 monophtongues du luxembourgeois produites par des locuteurs de la région centrale du Grand-Duché de Luxembourg. Le but de l'article était d'étendre les descriptions acoustiques sur les monophtongues du luxembourgeois existantes dans la littérature. Nous avons analysé les valeurs des trois premiers formants et de la durée : les analyses ont permis (i) d'obtenir des valeurs de référence pour les formants des monophtongues en position tonique dans la parole lue, (ii) de vérifier si les voyelles longues et brèves se distinguent tant par la durée acoustique que par le timbre, (iii) d'observer les

mouvements formantiques de la voyelle semi-ouverte [ɛː] avant /r/. Dans un premiers temps, les analyses ont tout d'abord dévoilé une importance de l'effet des formants et de la durée pour la distinction entre les paires de voyelles contrastives longues et brèves, appuyé par des tests statistiques appliqués à chaque paire. Ensuite, elles ont montré que le timbre, et notamment F3, s'est avéré particulièrement nécessaire pour la distinction entre [iː]-[i], [eː]-[e] et [aː]-[ɑ]. La première hypothèse, selon laquelle les paires de voyelles contrastives s'opposent à travers la durée et la qualité vocalique, serait alors confirmée pour nos données. Les résultats sont par contre valables uniquement pour les voyelles en position tonique dans des mots en contexte isolé. Par la suite, il est nécessaire de poursuivre des analyses plus approfondies, en les appliquant par exemple à des données dans un autre registre de parole. Dans un deuxième temps, nous avons observé que le [ɛː] se caractérise par un mouvement dynamique, semblable à celui de [ɛə] de l'anglais australien (WATSON ET HARRINGTON, 1999). La deuxième hypothèse doit donc être confirmée par une enquête plus approfondie sur la réalisation de [ɛː], par exemple en comparant le mouvement de cette voyelle à celui des autres monophtongues avant /r/ et à celui des diphtongues du luxembourgeois. Des analyses supplémentaires pourraient tenir compte d'autres facteurs, comme la position syllabique et la parole spontanée. Une comparaison avec des données sur des variétés de l'allemand permettrait également de mieux comprendre la fonction des caractéristiques des voyelles soulevées dans cet article, puisque les voyelles de l'allemand possèdent également des caractéristiques spécifiques telles que l'opposition entre les voyelles longues et brèves (PIROT *ET AL.*, 2015).

Comme le luxembourgeois a encore été peu étudié sur le plan linguistique et phonétique, notre travail représente une contribution à la recherche scientifique sur une variété de cette langue. Les perspectives de recherche sur le luxembourgeois sont grandissantes, tant en phonétique segmentale que suprasegmentale. Les possibilités d'études s'étendent vers plusieurs champs, tels que la sociophonétique, l'acquisition et la reconnaissance automatique de la parole. Notre démarche analytique nous a permis d'obtenir des résultats qui pourront servir de référence pour de futurs travaux descriptifs et comparatifs.

Références

ADDA-DECKER M., LAMEL L., ADDA G. (2014). Speech alignment and recognition experiments for Luxembourgish, *International Workshop on Spoken Language Technologies for Under-resourced Languages*, 53-60.

BOERSMA P., WEENINK D., (2014). *PRAAT: doing phonetics by computer [Computer Program], version 5.3.84.*

DELATTRE P., LIBERMAN A., COOPER, F., GERSTMAN, L. (1952). An experimental study of the acoustic determinants of vowel color: observations on one- and two-formant vowels synthesized from spectrographic patterns, *Word,* 195-210.

FEHLEN F., (2009). BaleineBis. Une enquête sur un marché linguistique multilingue en profonde mutation/Luxemburgs Sprachenmarkt im Wandel, *RED 12.*

GENDROT C., (2013). Réalisation et perception du /R/ standard français en finale de mot, *JCJC SHS 2 – Développement humain et cognition, langue et communication*.

GENDROT C., (2013). De la normalisation formantique des voyelles, *Méthodes et outils pour l'analyse phonétique des grands corpus oraux*, Cachan : Hermes/Lavoisier.

GEORGETON L., PAILLEREAU N., LANDRON S., GAO J., KAMIYAMA T., (2012). Analyse formantique des voyelles orales du français en contexte isolé : à la recherché d'une référence pour les apprenants de FLE, *Proceedings of the Joint Conference JEP-TALN-RECITAL* 1, Grenoble : ATALA/AFCP.

GILLES P., (1999). *Dialektausgleich im Lëtzebuergeschen*, Tübingen : Niemeyer.

GILLES P., TROUVAIN J., (2013). Luxembourgish, *Journal of the International Phonetic Association* 43, 67-74.

GOUDAILLIER J.-P., (1981). *Phonologie fonctionnelle et phonétique expérimentale : exemples empruntés au luxembourgeois*, Hambourg : Helmut Buske.

KEISER-BESCH D., (1976). Etude descriptive et analytique du vocalisme luxembourgeois, *Bulletin de Linguistique, Ethnologique et Toponymique* 20, 91-100.

KOHLER K., (1977). *Einführung in die Phonetik des Deutschen*, Berlin : Erich Schmidt Verlag.

NISHIDE K., (2014). Das Vokalsystem des Zentralluxemburgischen, *Neue Beiträge zur Germanistik* 13, 278-295.

PIROT G., SKUPINSKI P., POMPINO-MARSCHALL B., (2015). Production of vowel contrasts in Northern Standard German and Austrian Standard German, *International Congress of Phonetic Sciences*, 1-5.

THOMAS E., KENDALL T., (2007). *NORM : The vowel normalization and plotting suite*. [Online Resource: http://ncslaap.lib.ncsu.edu/tools/norm/]

TRAUNMÜLLER H., (1990). Analytical expressions for the tonotopic sensory scale, *Journal of the Acoustical Society of America* 88, 97-100.

TROUVAIN J., GILLES P., (2009). PhonLaF – phonetic online material for Luxembourgish as a foreign language, *Phonetics Teaching and Learning Conference*, 74-77.

VAISSIÈRE J., (2011). On the acoustic and perceptual characterization of reference vowels in a cross-language perspective, *International Congress of Phonetic Sciences*, 52-59.

WATSON C., HARRINGTON, J., (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels, *Journal of the Acoustical Society of America*, 458-468.

Auto-encodeurs pour la compréhension de documents parlés

Killian Janod[1, 3] Mohamed Morchid[1] Richard Dufour[1]
Georges Linarès[1] Renato De Mori[1, 2]
(1) LIA, 339 chemin des Meinajaries, Agroparc BP 1228, 84911 Avignon cedex 9, France
(2) McGill University, 845 Sherbrooke Street West, Montreal, Quebec, Canada H3A 0G4
(3) Orkis, 610 Rue Georges Claude Pôle d'activités d'Aix en Provence,13852 Aix-en-Provence, France
prénom.nom@univ-avignon.fr[1], rdemori@cs.mcgill.ca[2], kjanod@orkis.fr[3]

RÉSUMÉ

Les représentations de documents au moyen d'approches à base de réseaux de neurones ont montré des améliorations significatives dans de nombreuses tâches du traitement du langage naturel. Dans le cadre d'applications réelles, où des conditions d'enregistrement difficiles peuvent être rencontrées, la transcription automatique de documents parlés peut générer un nombre de mots mal transcrits important. Cet article propose une représentation des documents parlés très bruités utilisant des caractéristiques apprises par un auto-encodeur profond supervisé. La méthode proposée s'appuie à la fois sur les documents bruités et leur équivalent propre annoté manuellement pour estimer une représentation plus robuste des documents bruités. Cette représentation est évaluée sur le corpus DECODA sur une tâche de classification thématique de conversations téléphoniques atteignant une précision de 83% avec un gain d'environ 6%.

ABSTRACT

Auto-encoders for Spoken Document Understanding

Document representations based on neural embedding frameworks have recently shown significant improvements in different natural Language processing tasks. In the context of real application framework, the automatic transcription of spoken documents may result in several word errors, especially when very noisy conditions are encountered. This paper proposes an original representation of highly imperfect spoken documents based on the bottleneck features from a Supervised Deep auto-encodeur that takes advantage of both noisy automatic and clean manual transcriptions to improve the robustness of the document representation in a noisy environment. Results obtained on the DECODA theme classification task of dialogues reach an accuracy of more than 83% with a significant gain of about 6%.

MOTS-CLÉS : auto-encodeur, débruitage, reconnaissance de la parole, réseaux de neurones.

KEYWORDS: auto-encoder, denoising, speech recognition, neural networks.

1 Introduction

La recherche en compréhension du langage est très active notamment dans les disciplines d'analyse conversationnelle et de la parole, et de détection de thématique comme le montrent (Tur & De Mori, 2011) et (Purver, 2011). Un des axes d'innovation est lié à la détection de thèmes dans des conversions téléphoniques (voir figure 1) notamment grâce aux nombreuses possibilités applicatives qui en

FIGURE 1 – Un dialogue du copus DECODA annoté par un agent comme un problème de *carte de transport* qui contient aussi les thèmes secondaires (*infraction* et *carte de transport*).

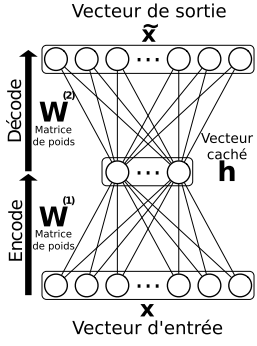


FIGURE 2 – Un auto-encodeur composé d'une couche d'entrée, une couche cachée et une couche de sortie. Pour des raisons de lisibilité les biais sont omis.

découlent. Les méthodes de cet axe s'appuient principalement sur les calculs de fréquence des mots transcrits. Dans un contexte téléphonique, la présence de différents locuteurs, environnements et médias de communication, génèrent une forte variabilité du signal. Dans un cadre automatique, cette variabilité implique que des erreurs de transcription soient commises par le système de reconnaissance automatique de la parole (SRAP) de façon non négligeable.

Cet article [1] propose une méthode où ces erreurs sont considérées comme du bruit perturbant les distributions de fréquence des mots. Cette approche s'appuie sur des auto-encodeurs avec débruitage (denoising autoencoders , DAE). Son but est de générer une distribution de fréquence des mots sans la perturbation générée par le bruit (Alain *et al.*, 2015). Cette méthode n'utilise pas de propriétés propres aux dialogues inter-humains ni à la reconnaissance automatique de la parole. Elle pourrait donc être *a priori* utile à tout type de ressources bruitées.

Les avancées récentes en apprentissage profond (LeCun *et al.*, 2015) ont montré d'excellentes performances dans des domaines applicatifs comme le traitement du langage (Yu *et al.*, 2010) et de la parole (Mohamed *et al.*, 2009). Dans ce cadre, les auto-encodeurs sont souvent utilisés pour obtenir des représentations latentes capables de capturer suffisamment d'informations pour reconstruire les données d'origine. Ces représentations sont utilisées habituellement comme pré-entraînement de réseaux de neurones profonds (*deep neural network* , DNN) (Erhan *et al.*, 2010). De nombreux DNN ont déjà été proposés à des fins de débruitage. Parmi eux, (Gallinari *et al.*, 1987) propose des "mémoires associatives" pour retrouver de l'information à partir de données partielles. Plus récemment, des solutions s'appuyant sur des méthodes d'apprentissage non-supervisé utilisant des données issues de conditions homogènes ont été proposées (Vincent *et al.*, 2008; LeCun *et al.*, 2015). Les auto-encodeurs permettant le débruitage (DAE) ont été proposés (Vincent *et al.*, 2008) pour améliorer la robustesse du processus de reconstruction en présence de données bruitées. Ces DAE ont prouvé leur intérêt dans de nombreux domaines, allant de la biologie (Camacho *et al.*, 2015) jusqu'au traitement de la musique (Sarroff & Casey, 2014). Ces DAE apprennent à reproduire un vecteur sain à partir du même vecteur artificiellement corrompu par un bruit additif. Ils sont efficaces quand les données d'entrée et de sortie possèdent des conditions homogènes et ne portent pas une information creuse (*sparse*). Durant le processus d'apprentissage, l'erreur à rétro-propager est calculée entre le

1. Ce travail a été réalisé dans le cadre du projet GaFes financé par l'Agence Nationale de la Recherche (ANR) sous le contrat ANR-14-CE24-0022.

document produit par le réseau et le document propre d'origine. Dans le cas de données naturellement bruitées, le type de bruit est inconnu et donc plus difficile à caractériser et à nettoyer.
Cet article propose une solution pour créer un document propre à partir d'une version corrompue sans connaissance *a priori* du bruit. Elle produit un ensemble de caractéristiques latentes robustes via un auto-encodeur profond supervisé (*deep bottlenecked autoencoder* , BDAE). Le BDAE tire profit à la fois des données annotées (*i.e.* propres) par un expert humain et des transcriptions produites par un SRAP. De ces données, une transformation non linéaire est apprise entre un espace dit "bruité" et un espace dit "propre" sans appliquer de fonction de corruption artificielle aux données.
La suite de l'article est organisée comme suit : l'approche choisie est détaillée dans la section 2, le protocole expérimental et les résultats étant présentés section 3 et 4. Enfin, la section 5 ouvre des perspectives de travail.

2 Approche proposée

Cette section présente les concepts de base des auto-encodeurs, en s'intéressant aux auto-encodeurs avec débruitage (DAE) et aux auto-encodeurs profonds supervisés (BDAE).

2.1 Description des auto-encodeurs

Les auto-encodeurs (AE) sont des réseaux de neurones simples composés de trois couches. La première couche et la couche cachée forment l'encodeur, la couche cachée ainsi que la dernière couche formant le décodeur comme décrit dans la figure 2. L'encodeur calcule, à partir de $\mathbf{x}$, le vecteur $\mathbf{h}$ de taille m (nombre de neurones cachés) ainsi : $\mathbf{h} = \sigma(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)})$
où $\mathbf{W}^1$ est une matrice de taille $m \times n$ et $\mathbf{b}^{(1)}$ un vecteur de biais de taille m. $\sigma(.)$ est une fonction d'activation de type tangente hyperbolique définie par : $\sigma(\mathbf{y}) = \frac{e^{\mathbf{y}} - e^{-\mathbf{y}}}{e^{\mathbf{y}} + e^{-\mathbf{y}}}$. Le décodeur cherche à reconstruire le vecteur $\mathbf{x}$ à partir de la couche cachée $\mathbf{h}$. Le résultat de cette reconstruction est le vecteur $\tilde{\mathbf{x}}$ tel que : $\tilde{\mathbf{x}} = \sigma(\mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)})$ où $\tilde{\mathbf{x}}$ est un vecteur de taille m, $\mathbf{W}^{(2)}$ est une matrice de poids de taille $n \times m$ et $\mathbf{b}^{(2)}$ est un vecteur de biais de taille n. Durant l'apprentissage, l'auto-encodeur tente de réduire une erreur de reconstruction l entre $\mathbf{x}$ et $\tilde{\mathbf{x}}$. Il utilise l'erreur quadratique moyenne (MSE) ($l_{\text{MSE}}(\mathbf{x}, \tilde{\mathbf{x}}) = ||\mathbf{x} - \tilde{\mathbf{x}}||^2$) de manière à minimiser l'erreur de reconstruction totale L_{MSE} avec l'ensemble de paramètres $\theta = \{\mathbf{W}^{(2)}, \mathbf{b}^{(1)}, \mathbf{W}^{(2)}, \mathbf{b}^{(2)}\}$:

$$L_{\text{MSE}}(\theta) = \frac{1}{d} \sum_{\mathbf{x} \in D} l_{\text{MSE}}(\mathbf{x}, \tilde{\mathbf{x}}) = \frac{1}{d} \sum_{\mathbf{x} \in D} ||\mathbf{x} - \tilde{\mathbf{x}}||^2 \quad (1)$$

Il est souvent fait référence aux capacités qu'ont les réseaux de neurones profonds à encoder une information d'un plus haut niveau d'abstraction au fur et à mesure des couches cachées successives (Bengio *et al.*, 2007; Hinton *et al.*, 2006). Dans un réseau de neurones empilés (*stacked autoencoder*, SAE) avec k couches cachées, les caractéristiques latentes de la i-ème couche, $\mathbf{h}^{(i)}$, pour un vecteur $\mathbf{x}$ donné, sont calculées comme suit : $\mathbf{h}^{(i)} = \sigma(\mathbf{W}^{(i)}\mathbf{h}^{(i-1)} + \mathbf{b}^{(i)})$ $\forall i \in \{1, \ldots, k\}$ et $\mathbf{h}^{(0)} = \mathbf{x}$. De plus, chaque couche est pré-entraînée comme le serait un auto-encodeur simple pour un nombre d'itérations défini. Le vecteur ainsi appris $\mathbf{h}^{(i)}$ est conservé et utilisé pour entraîner la couche suivante $\mathbf{h}^{(i+1)}$. Ce pré-entraînement dit "gourmand" est réalisé progressivement en commençant par $\mathbf{h}^{(0)}$ jusqu'à obtention de la niveaux d'abstraction recherché. L'objectif d'un auto-encodeur est d'encoder

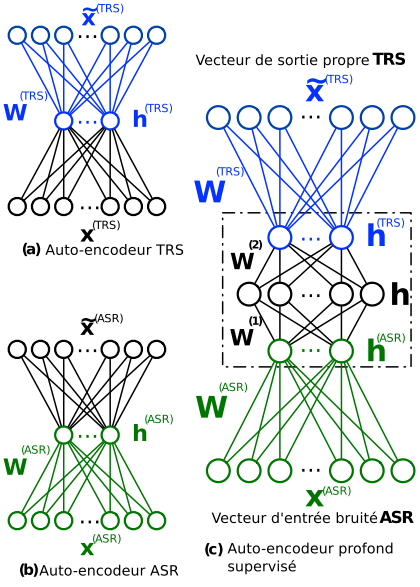


FIGURE 3 – Réseau proposé (c) initialisé avec les poids venant des AE ASR (a) et TRS (b).

puis décoder un vecteur $\mathbf{x}$ vers/depuis un espace latent $\mathbf{h}$. Pendant le processus d'apprentissage, les auto-encodeurs ne peuvent pas toujours séparer l'information pertinente du bruit résiduel pour une distribution donnée. Pour cette raison, (Vincent *et al.*, 2008) propose le DAE qui corrompt artificiellement les vecteurs d'entrée.
La figure 4 montre le schéma du fonctionnement d'un DAE. Dans ce réseau, le vecteur d'entrée $\mathbf{x}$ est considéré comme "propre". Le DAE vise à obtenir une reconstruction propre robuste à la corruption. Pour y arriver, $\mathbf{x}$ est corrompu artificiellement par une fonction de bruitage aléatoire (Vincent *et al.*, 2008). Cette entrée corrompue $\mathbf{x}^{(\text{corrompu})}$ est ensuite projetée dans une couche cachée $\mathbf{h} = f_{(\mathbf{W}^{(1)},\mathbf{b}^{(1)})} = \sigma(\mathbf{W}^{(1)}\mathbf{x}^{(\text{corrompu})} + \mathbf{b}^{(1)})$. Le vecteur $\tilde{\mathbf{x}}$ est reconstruit de manière à minimiser l'erreur de reconstruction $L(\mathbf{x}, \tilde{\mathbf{x}})$.

La motivation de ce type de réseaux de neurones est qu'une bonne représentation $\mathbf{h}$ d'un vecteur d'entrée $\mathbf{x}$ est invariante aux perturbations que peut appliquer un bruit à $\mathbf{x}$. La distribution conditionnelle utilisée pour générer $\mathbf{x}^{(\text{corrompu})}$ est induite par apprentissage. Le problème abordé dans cet article est différent, les caractéristiques étant extraites depuis des données provenant de conversations téléphoniques enregistrées avec du bruit de fond difficile à prédire et à caractériser.

2.2 Génération de caractéristiques robustes au bruit

Dans cet article, nous voulons obtenir une représentation d'un document, issue d'un SRAP, qui soit robuste aux erreurs et efficace pour la détection de thèmes. Cette représentation robuste s'appuie sur des caractéristiques issues d'un BDAE reposant sur une transcription automatique et manuelle. Cette architecture (voir figure 3-c) utilise les vecteurs imparfaits issus du SRAP (ASR) et les vecteurs transcrits manuellement (TRS). L'estimation des paramètres est réalisée en utilisant les architectures décrites dans les figures 3-a et -b.

Initialisation : Comme noté dans (Hinton *et al.*, 2006), l'initialisation des paramètres d'estimation dans l'architecture profonde est critique. Dans cette optique, les matrices de poids du BDAE ($\mathbf{W}^{(\text{ASR})}$ et $\mathbf{W}^{(\text{TRS})}$) sont initialisées à partir des poids appris par des auto-encodeurs classiques, comme le montrent les figures 3.
Processus d'apprentissage : Un nouvel entraînement (pointillé dans figure 3-c) est réalisé pour estimer une transformation non-linéaire entre l'espace latent bruité $\mathbf{h}^{(\text{ASR})}$ (en vert) et l'espace latent propre $\mathbf{h}^{(\text{TRS})}$ (en bleu) en passant par une couche cachée intermédiaire. L'erreur de reconstruction totale L_{MSE} définie dans l'équation 1 est calculée avec une erreur l_{MSE} entre le vecteur de sortie $\tilde{\mathbf{x}}^{(\text{TRS})}$ et le document propre $\mathbf{x}^{(\text{TRS})}$:

$$l_{\text{MSE}}(\mathbf{h}^{(\text{TRS})}, \tilde{\mathbf{h}}^{(\text{TRS})}) = ||\mathbf{h}^{(\text{TRS})} - \tilde{\mathbf{h}}^{(\text{TRS})}||^2 \quad (2)$$

Le processus d'extraction des caractéristiques, pour un document bruité donné $\mathbf{x}^{(\text{ASR})}$, nécessite une étape d'encodage puis de décodage décrites ci-dessous :
Phase d'encodage : un vecteur d'entrée $\mathbf{x}^{(\text{ASR})}$ est projeté dans un espace latent bruité pour obtenir un vecteur $\mathbf{h}^{(\text{ASR})}$, puis $\mathbf{h}^{(\text{ASR})}$ est projeté dans un espace intermédiaire pour obtenir $\mathbf{h}$;
Phase de décodage : le vecteur $\mathbf{h}$ est ensuite projeté dans l'espace latent propre pour générer $\mathbf{h}^{(\text{TRS})}$ qui permettra de reconstruire le vecteur $\tilde{\mathbf{x}}^{(\text{TRS})}$.

3 Protocole Expérimental

La robustesse de la représentation fondée sur le BDAE est évaluée dans un cadre de détection de thèmes sur le corpus DECODA (Bechet *et al.*, 2012). Le corpus DECODA (Bechet *et al.*, 2012) est un ensemble de conversations téléphoniques provenant du service de gestion clientèle de la RATP. Ce corpus est utilisé pour réaliser des tâches de détection de thèmes dans ces conversations. Il est composé de 1 242 conversations (740 pour l'apprentissage, 175 pour le développement et 327 pour le test) correspondant à 74 heures de signal et découpé en 8 catégories thématiques : *Itinéraire, Objets trouvés, Horaire, Carte de transport, État du trafic, Prix du ticket, Infractions, Offres spéciales*. Les conversations ont été transcrites et annotées manuellement.
Un système de reconnaissance automatique de la parole (SRAP) est utilisé pour transcrire automatiquement le corpus DECODA. Il utilise un modèle acoustique triphone de 230 000 gaussiennes appris sur 150 heures de parole dans des conditions téléphoniques, ainsi qu'un modèle de langue 3-grammes spécifique de 5 782 mots. Le taux d'erreur-mots (*word error rate, WER)* est de 33,8 % sur les données d'entraînement, 45,2 % sur le développement, et 49,5 % sur le test. Ces WER élevés sont principalement dus aux mauvaises conditions acoustiques et aux disfluences verbales. Des WER proches ont été rapportés dans des conditions similaires (Garnier-Rizet *et al.*, 2008).
Un sous-ensemble de mots discriminants via le produit entre l'inverse de la fréquence inter-documents et la pureté du mot définie par le critère de Gini (IDF.G) est construit à partir du corpus d'apprentissage. Pour chaque thème du corpus, 100 mots spécifiques sont identifiés, formant un vocabulaire de 707 mots. Pour un corpus D donné, un vecteur de caractéristiques x est défini par les éléments $\mathbf{x}_i$ calculés ainsi : $\mathbf{x}_i = |t_i| \times \Delta(t_i)$où $|t_i|$ est le nombre d'occurrences du i-ème mot dans le document et $\Delta(t_i)$ est le IDF.G. Une classification thématique des documents est réalisée par un Perceptron multi-couches (*Multi-layer Perceptron*, MLP).

Deux auto-encodeurs AE_{ASR} et AE_{TRS} utilisant les caractéristiques $\mathbf{x}^{(\text{TRS})}$ et $\mathbf{x}^{(\text{ASR})}$, et avec chacun une couche cachée de 50 neurones artificiels ($\mathbf{h}^{(\text{ASR})}$ ou $\mathbf{h}^{(\text{TRS})}$) sont entraînés. Un auto-encodeur profond supervisé BDAE (voir figure 3-c) est entraîné pour extraire les caractéristiques latentes des

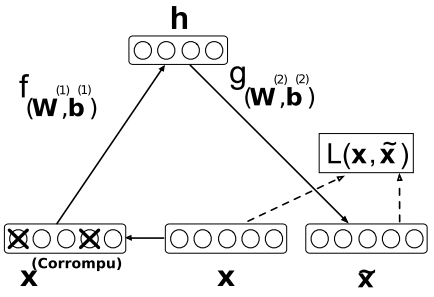


FIGURE 4 – Un auto-encodeur avec débruitage ayant une entrée naturellement bruité $\mathbf{x}^{(\text{ASR})}$ et la production propre voulue $\mathbf{x}^{(\text{TRS})}$. L'erreur de reconstruction L est évaluée entre la sortie observée $\tilde{\mathbf{x}}^{(\text{TRS})}$ et $\mathbf{x}^{(\text{TRS})}$.

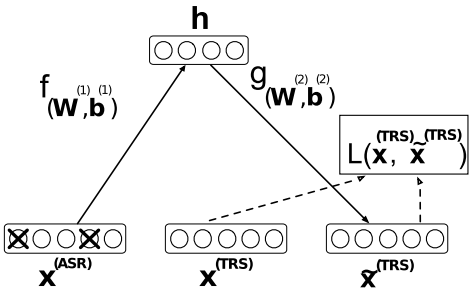


FIGURE 5 – Un auto-encodeur avec débruitage ayant une entrée corrompue $\mathbf{x}^{(\text{ASR})}$ et la production propre voulue $\mathbf{x}^{(\text{TRS})}$. L'erreur de reconstruction L est évaluée entre la sortie observée $\tilde{\mathbf{x}}^{(\text{TRS})}$ et $\mathbf{x}^{(\text{TRS})}$.

documents dans une couche cachée centrale **h** de 300 neurones. Il utilise des vecteurs de fréquence des mots bruités $\mathbf{x}^{(\text{ASR})}$ comme vecteurs d'entrée et de sortie $\tilde{\mathbf{x}}^{(\text{TRS})}$. Le premier et le dernier vecteur caché $\mathbf{h}^{(\text{ASR})}$ et $\mathbf{h}^{(\text{TRS})}$ ont leurs matrices de poids, $\mathbf{W}^{(\text{ASR})}$ et $\mathbf{W}^{(\text{TRS})}$, initialisées à partir de AE_{ASR} et AE_{TRS} sans réapprentissage ensuite.

À des fins de comparaison, 4 réseaux de neurones artificiels supplémentaires ont été évalués : (1) un auto-encodeur empilé **SAE** avec les entrées provenant des documents ASR et $k = 4$ couches cachées de taille successive 50, 300, 50 et 707 (miroir du BDAE) ; (2) un auto-encodeur profond supervisé avec apprentissage global et sans pré-apprentissage appelé **FBDAE** ; (3) un auto-encodeur avec débruitage **DAE** dont les vecteurs d'entrée proviennent des documents ASR et la sortie des documents TRS avec une couche cachée de taille 50, en miroir de la figure 5 ; (4) un auto-encodeur avec débruitage profond **DDAE** avec 3 couches cachées identiques au BDAE.

Les différentes architectures sont évaluées sur la tâche d'identification thématique (voir tableaux 1 et 2). Les hypothèses thématiques sont émises par un MLP avec 256 neurones et une couche de sortie composée de 8 neurones correspondants aux 8 thèmes de la tâche. L'apprentissage est réalisé sur une carte graphique Nvidia GeForce GTX TITAN X. L'apprentissage du MLP nécessite environ 8 minutes de calcul. Pour les auto-encodeurs, les temps suivants sont rapportés : AE simples 10 minutes ; DAE, DDAE et DBAE 25 minutes ; et enfin 50 minutes pour le SAE.

4 Expériences et Résultats

Les résultats pour la classification de thèmes avec les auto-encodeurs simples et profonds sont répertoriés dans les sections 4.1 et 4.2. Enfin, les avantages du BDAE sont discutés section 4.3.

4.1 Les Auto-Encodeurs Simples

Le tableau 1 présente les précisions de classification obtenues avec les caractéristiques générées par les auto-encodeurs à la fois avec la transcription automatique (ASR) et avec la transcription manuelle (TRS). Pour comparaison, les précisions obtenues par les MLP sont rapportées pour les

Méthode	Entrée donnée	Sortie	Précision sur le test		
			x	couche cachée **h**	sortie $\tilde{\mathbf{x}}$
AE_{ASR}	ASR	ASR	77.1	**81**	79
AE_{TRS}	TRS	TRS	83.4	84.1	83.7
DAE	ASR	TRS	77.1	74.3	70.3

TABLE 1 – Précisions de classification (%) avec les caractéristiques produites par les auto-encodeurs (Figure 3-a et Figure 3-b) et un auto-encodeur avec débruitage mono-couche (Figure 5).

différentes couches des AE (**x**, **h** et $\tilde{\mathbf{x}}$) utilisées comme vecteurs d'entrée. Dans ces conditions, nous remarquons que les meilleures précisions sont obtenues avec des couches cachées apprises dans des conditions homogènes (ASR $\rightarrow$ ASR et TRS $\rightarrow$ TRS) apportant un gain de $3,9$ et $0,7$ points pour les documents ASR et TRS respectivement. La précision obtenue avec le DAE ASR $\rightarrow$ TRS est largement détériorée dès que la représentation gagne un niveau d'abstraction. Comme attendu, l'erreur contenue dans les transcriptions automatiques de documents bruités fait diminuer la précision de la classification, de $84,1$ % à 81 % pour le vecteur caché **h** par exemple. La précision de $84,1\%$ obtenue avec l'auto-encodeur sur les documents transcrits manuellement représente la borne supérieure visée dans le cas où notre méthode réaliserait un débruitage optimal.

4.2 Les Auto-Encodeurs Profonds

Le tableau 3 compare la précision de classification des caractéristiques issues de différents auto-encodeurs profonds classiques et le BDAE proposé utilisant les documents ASR en entrée. Cette table ne montre pas les résultats pour les vecteurs **x** et les couches cachées du BDAE qui sont déjà présentées dans le tableau 1 (*e.g.* : $\mathbf{h}(\text{ASR})$ de BDAE $= \mathbf{h}$ de AE_{ASR}). Les meilleurs résultats sont obtenus avec le BDAE avec un score notable de $83,2$ %. Ce résultat est proche des performances obtenues utilisant des caractéristiques extraites de l'auto-encodeur sur les documents propres (AE_{TRS}). L'intuition initiale sur le réapprentissage du BDAE est vérifiée. En effet, la précision obtenue avec l'utilisation du FBDAE est détériorée atteignant $76,5$ %. Simplement augmenter le niveau d'abstraction n'améliore pas non plus la robustesse au bruit et fait chuter aussi la précision à $69,4$ % avec la couche $\mathbf{h}^{(3)}$ du DDAE. L'auto-encodeur avec débruitage empilé (SAE) obtient de bons résultats avec un gain de $9,5$ et $5,5$ points comparativement aux DDAE et FBDAE. La qualité des résultats du SAE s'explique principalement par le fait que ce réseau est entraîné uniquement avec les données issues des documents bruités ASR, les erreurs impliquées par la reconstruction d'un document dans un espace différent ne sont pas rétro-propagées à travers ces couches cachées.

4.3 Discussions

Le tableau 2 présente les meilleures précisions des différentes architectures comparées dans cet article. En premier lieu, BDAE exclu, les AE_{ASR} et SAE sont les méthodes les plus robustes. Ces résultats s'expliquent par le fait que ces deux méthodes sont capables de supprimer une importante partie du bruit contenu dans les documents. Par contre, ces deux méthodes n'arrivent pas à s'approcher des résultats sur les documents propres. Le tableau 2 montre que AE_{ASR} est aussi capable de retirer un bruit du document mais avec une efficacité bien moindre.
La précision avec les caractéristiques extraites du BDAE approche les $83,2$ %, seulement $0,9$ point sous la précision obtenue avec les documents propres (TRS) (voir tableau 1). Ce résultat montre

Methode employée	Vecteurs	Test Précision
DDAE	$\mathbf{h}^{(1)}$	72.5
DAE	$\mathbf{h}$	74.3
FBDAE	$\mathbf{h}$	76.5
TF.IDF.G	–	77.1
AE_{ASR}	$\mathbf{h}$	81
SAE	$\mathbf{h}^{(3)}$	82.0
BDAE proposé	$\mathbf{h}$	**83.2**

TABLE 2 – Meilleures précisions de classification (%) observées sur les caractéristiques extraites des documents ASR

auto-encodeur employé	Entrée	Sortie	couche vecteur	Test Précision
Autoencodeur	ASR	–	$\mathbf{h}^{(1)}$	81.7
Profonds	ASR	–	$\mathbf{h}^{(2)}$	82.0
Empilé	ASR	–	$\mathbf{h}^{(3)}$	80.1
(**SAE**)	ASR	–	$\mathbf{h}^{(4)}$	81.0
Autoencodeur	ASR	TRS	$\mathbf{h}^{(1)}$	72.5
Débruitant	ASR	TRS	$\mathbf{h}^{(2)}$	70
Profonds	ASR	TRS	$\mathbf{h}^{(3)}$	69.4
(**DDAE**)	ASR	TRS	$\tilde{\mathbf{x}}$	69.7
Autoencodeur	ASR	TRS	$\mathbf{h}^{(\text{ASR})}$	69.7
Profond	ASR	TRS	$\mathbf{h}$	76.5
Réappris	ASR	TRS	$\mathbf{h}^{(\text{TRS})}$	73.4
(**FBDAE**)	ASR	TRS	$\mathbf{h}$	71.9
BDAE proposé	ASR	TRS	$\mathbf{h}$	**83.2**

TABLE 3 – Précision de classification (%) avec des vecteurs issus de différentes configurations.

qu'un faible pourcentage des erreurs de reconstruction affecte les performances de classification des documents transcrits automatiquement.
Enfin, les faibles résultats montrés dans le tableau 2 des réseaux DDAE, DAE, FBDAE montrent bien que tenter de supprimer l'ensemble du bruit en une fois est une mauvaise idée. Le bruit dans les documents ASR est trop complexe pour être supprimé directement. Avec le réseau BDAE proposé, les première et dernière couches capitalisent sur les capacités de AE_{ASR} et AE_{TRS} pour supprimer un bruit résiduel. Ensuite, la couche cachée de transfert peut se concentrer sur un bruit plus complexe. Cette méthode permet à ce réseau de produire une représentation plus propre et robuste.

5 Conclusion

Cet article propose une représentation originale des documents fondée sur des caractéristiques réduites provenant d'un auto-encodeur profond supervisé. Cette représentation est appliquée à un problème d'identification de thèmes dans des documents transcrits automatiquement. Les matrices de poids de ce réseau de neurones profond sont extraites de deux auto-encodeurs classiques. Ces deux auto-encodeurs sont entrainés sur des documents corrompus pour le premier (ASR), et des documents propres pour le second (TRS). Ensuite une transformation non-linéaire des documents ASR vers les documents TRS est apprise indépendamment. Ainsi, le système préserve la projection des documents corrompus vers la représentation latente corrompue et la projection des documents propres vers la représentation latente propre. L'architecture proposée permet un gain de plus de $6,7$ points dans la projection latente réduite comparé à l'homologue réappris. Un écart de seulement 0.9 point avec les documents annotés manuellement a alors été observé. Les prochains travaux de cette étude préliminaire viseront à prendre en compte la structure des documents en remplaçant les couches simples dans le BDAE par des couches récurrentes pour utiliser les propriétés des réseaux de neurones récurrents tels que les *Long-Short Term Memory (LSTM) autoencoder (Cho* et al.*, 2014)* ou les *Gated Recurrent units* (Droniou & Sigaud, 2013). En effet, prendre en compte l'information structurelle peut apporter des informations supplémentaires sur le bruit mais aussi permettre de reconstruire des documents complets en plus d'une représentation latente.

Références

ALAIN G., BENGIO Y., YAO L., YOSINSKI J., THIBODEAU-LAUFER E., ZHANG S. & VINCENT P. (2015). Gsns : Generative stochastic networks. *arXiv preprint arXiv :1503.05571*.

BECHET F., MAZA B., BIGOUROUX N., BAZILLON T., EL-BEZE M., DE MORI R. & ARBILLOT E. (2012). : LREC'12.

BENGIO Y., LECUN Y. *et al.* (2007). Scaling learning algorithms towards ai. *Large-scale kernel machines*, **34**(5).

CAMACHO F., TORRES R. & RAMOS-POLLÁN R. (2015). Feature learning using stacked autoencoders to predict the activity of antimicrobial peptides. In *Computational Methods in Systems Biology*, p. 121–132 : Springer.

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*.

DRONIOU A. & SIGAUD O. (2013). Gated autoencoders with tied input weights. In *International Conference on Machine Learning*, p.x̃.

ERHAN D., BENGIO Y., COURVILLE A., MANZAGOL P.-A., VINCENT P. & BENGIO S. (2010). Why does unsupervised pre-training help deep learning ? *The Journal of Machine Learning Research*, **11**, 625–660.

GALLINARI P., LECUN Y., THIRIA S. & FOGELMAN-SOULIE F. (1987). Memoires associatives distribuees. *Proceedings of COGNITIVA*, **87**, 93.

GARNIER-RIZET M., ADDA G., CAILLIAU F., GAUVAIN J., GUILLEMIN-LANNE S., LAMEL L., VANNI S. & WAAST-RICHARD C. (2008). Callsurf-automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content. In *Proceedings of LREC*.

HINTON G. E., OSINDERO S. & TEH Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, **18**(7), 1527–1554.

LECUN Y., BENGIO Y. & HINTON G. (2015). Deep learning. *Nature*, **521**(7553), 436–444.

MOHAMED A., DAHL G. & HINTON G. (2009). Deep belief networks for phone recognition,[in :] nips workshop on deep learning for speech recognition and related applications.

PURVER M. (2011). Topic segmentation. *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*, p. 291–317.

SARROFF A. M. & CASEY M. (2014). Musical audio synthesis using autoencoding neural nets.

TUR G. & DE MORI R. (2011). *Spoken language understanding : Systems for extracting semantic information from speech*. John Wiley & Sons.

VINCENT P., LAROCHELLE H., BENGIO Y. & MANZAGOL P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, p. 1096–1103 : ACM.

YU D., WANG S., KARAM Z. & DENG L. (2010). Language recognition using deep-structured conditional random fields. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, p. 5030–5033 : IEEE.

Autoapprentissage pour le regroupement en locuteurs : premières investigations

Gaël Le Lan[1,2] Sylvain Meignier[1] Delphine Charlet[2] Anthony Larcher[1]
(1) LIUM, Université du Maine, France
(2) Orange Labs, France
prenom.nom@lium.univ-lemans.fr, prenom.nom@orange.com

RÉSUMÉ

Cet article traite de l'autoapprentissage d'un système *i-vector/PLDA* pour le regroupement en locuteurs de collections d'archives audiovisuelles françaises. Les paramètres d'extraction des i-vectors et du calcul des scores PLDA sont appris de façon non supervisée sur les données de la collection elle-même. Différents mélanges de données cibles et de données externes sont comparés pour la phase d'apprentissage. Les résultats expérimentaux sur deux corpora cibles distincts montrent que l'utilisation des données des corpora en question pour l'apprentissage itératif non supervisé et l'adaptation des paramètres de la PLDA peut améliorer un système existant, appris sur des données annotées externes. De tels résultats indiquent que la structuration automatique en locuteurs de petites collections non annotées ne devrait reposer que sur l'existence d'un corpus externe annoté, qui peut être spécifiquement adapté à chaque collection cible. Nous montrons également qu'une collection suffisamment grande peut se passer de l'utilisation de ce corpus externe.

ABSTRACT

First investigations on self trained speaker diarization

This paper investigates self trained cross-show speaker diarization applied to collections of French TV archives, based on an *i-vector/PLDA* framework. The parameters used for i-vectors extraction and PLDA scoring are trained in a unsupervised way, using the data of the collection itself. Performances are compared, using combinations of target data and external data for training. The experimental results on two distinct target corpora show that using data from the corpora themselves to perform unsupervised iterative training and domain adaptation of PLDA parameters can improve an existing system, trained on external annotated data. Such results indicate that performing speaker indexation on small collections of unlabeled audio archives should only rely on the availability of a sufficient external corpus, which can be specifically adapted to every target collection. We show that a minimum collection size is required to exclude the use of such an external bootstrap.

1 Introduction

La tâche de Segmentation et Regroupement en Locuteurs (SRL) vise à étiqueter les locuteurs dans un ou plusieurs enregistrements audios, sans connaissance a priori des locuteurs. L'augmentation constante des volumes de données nécessite une indexation efficace. La SRL appliquée à des collections est une tâche globale qui consiste à traiter un ensemble d'enregistrements audios bruts (non segmentés en tours de parole) afin d'identifier de manière unique les tours de paroles de chaque locuteur. Cette tâche se décompose généralement en deux étapes : la SRL intra-enregistrement, où il

s'agit de segmenter et regrouper les occurrences des locuteurs au sein d'un même enregistrement, et le regroupement inter-enregistrements, qui vise à regrouper les locuteurs intra-enregistrements (Meignier *et al.*, 2002). D'autres implémentations sont possibles, où tous les enregistrements peuvent être concaténés en un super-enregistrement, pour ensuite être traités comme un problème artificiel de SRL intra-enregistrement (Tran *et al.*, 2011). L'approche en deux étapes, plus naturelle, est la plus couramment utilisée. Elle permet de traiter de grands volumes de données comme des archives audiovisuelles ou radiophoniques (Le Lan *et al.*, 2016; Dupuy *et al.*, 2014a; Yang *et al.*, 2011; Tran *et al.*, 2011; Van Leeuwen, Proc Odyssey 2010), des enregistrements téléphoniques (Shum *et al.*, Proc Odyssey 2014, 2013; Karam & Campbell, 2013; Ghaemmaghami *et al.*, 2012), ou encore d'enregistrements de réunions (Ferràs & Bourlard, 2012).

L'état de l'art en reconnaissance du locuteur repose sur le système *i-vector/PLDA*, qui requiert des données annotées en locuteurs, indiquant l'identité et les tours de parole de ceux-ci. La modélisation de la variabilité inter-locuteur est une étape clé. Pour la réaliser, le corpus d'apprentissage doit contenir plusieurs occurrences d'un même locuteur dans des conditions acoustiques variées. Les annotations manuelles en locuteurs sont coûteuses et rarement disponibles. Ainsi, deux approches peuvent être utilisées pour entraîner un système automatique : 1) un apprentissage non supervisé sur les données cibles, 2) une approche supervisée sur des données d'entraînement annotées en locuteurs. Dans cette deuxième approche, la différence de conditions acoustiques entre le corpus d'apprentissage et le corpus cible induit une dégradation des performances sur la collection cible. Des solutions de compensation ont déjà été proposées pour la tâche de vérification du locuteur, basées sur l'adaptation au domaine de manière non supervisée (Shum *et al.*, Proc Odyssey 2014). Dans le contexte de l'apprentissage non supervisé d'un modèle PLDA [1], la notion d'apprentissage itératif (Liu *et al.*, 2014) a été proposée pour améliorer la qualité des modèles. Dans les deux articles précités, l'apprentissage était fait sur des enregistrements mono locuteur dont l'identité est inconnue. Dans (Le Lan *et al.*, 2016), nous avons montré qu'un système de SRL pour une collection, entraîné de façon non supervisée sur des enregistrements multi locuteurs non segmentés en tour de parole, pouvait être aussi performant qu'un système appris de manière supervisée sur les mêmes données. Ces expériences nous ont montré que les annotations d'un corpus d'apprentissage ne sont pas obligatoires.

Cette étude porte sur l'autoapprentissage des modèles *i-vector/PLDA* utilisés lors des phases de regroupement intra-et inter-enregistrements à partir des enregistrements eux-mêmes, sans l'utilisation de données d'apprentissage ou d'adaptation externes. L'idée est de proposer un système de SRL utilisant le minimum de connaissances a priori. Dans ce contexte, nous étudions plusieurs variantes du système de regroupement *i-vector/PLDA*, appris avec ou sans données externes. Nous évaluons aussi des modèles appris à partir de la combinaison de données étiquetées et non étiquetées en locuteur. Nous commençons par décrire le système de SRL en précisant le périmètre de l'autoapprentissage pour la tâche visée. Ensuite, nous détaillons les données utilisées pour les expériences avant de conclure avec l'étude des performances du système proposé et des améliorations possibles.

2 Systèmes de SRL intra-et inter-enregistrements

La figure 1 décrit les traitements pour l'étape de SLR inter-enregistrements et l'étape intra-enregistrement résumés dans les deux paragraphes ci-dessous. Nous invitons le lecteur à se reporter à (Le Lan *et al.*, 2016) pour une description complète du processus de SRL.

1. PLDA : Probabilistic Linear Discriminant Analysis

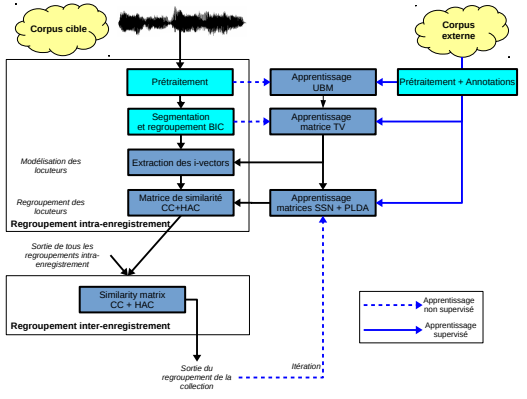


FIGURE 1 – Vue d'ensemble du système de SRL, entraîné de façon supervisé (traits bleus continus), non supervisé (traits pointillés bleus) et semi-supervisé (tous les traits bleus).

SRL intra-enregistrement : après l'extraction de 13 MFCC et une étape de détection de la parole, un système classique de segmentation et de classification BIC (Barras *et al.*, 2006) permet d'obtenir des classes très pures contenant les tours de parole d'un locuteur. À ce stade, nous disposons de suffisamment de données pour apprendre à partir de MFCC normalisés (moyenne et variance) et étendus avec leurs dérivés premières et secondes des modèles plus complexes comme les *i-vectors* (Kenny, 2010). Après leur apprentissage, ces modèles sont normalisés par la méthode décrite dans (Bousquet *et al.*, 2012).

Un rapport de vraisemblance entre chaque paire de *i-vectors* est calculé en utilisant un modèle PLDA (dimension 200 pour la matrice *inter-classes* et sans matrice *intra-classes*). Ces paires forment une matrice de scores sur laquelle nous appliquons l'algorithme de regroupement $CC + HAC$ à base de graphe et de regroupement hiérarchique décrit dans (Dupuy *et al.*, 2014b).

SRL inter-enregistrements : après avoir traité chaque émission séparément, la collection est considérée dans son ensemble et le même algorithme $CC + HAC$ est appliqué afin de regrouper les classes obtenues dans chaque émission, chacune étant représentée par un unique *i-vector*.

3 Autoapprentissage du système automatique

Après avoir montré dans (Le Lan *et al.*, 2016) que, pour un corpus de grande dimension, l'apprentissage non supervisé d'un modèle de SRL pouvait donner des performances comparables à un apprentissage supervisé, nous étudions ici le cas où aucun corpus externe (annoté ou non) n'est disponible pour l'apprentissage et où les seules données d'apprentissage sont issues du corpus cible lui-même, qui est non annoté.

La figure 1 représente les processus d'apprentissage d'un système de SRL supervisé (traits continus bleus) et non supervisé (traits pointillés bleus). Ce dernier est similaire à celui décrit dans (Le Lan *et al.*, 2016) mais ici, le corpus cible sert de corpus d'apprentissage. Nous comparons cette approche à un système de référence, appris de façon classique sur un corpus d'apprentissage annoté et à un système oracle appris sur le corpus cible avec des annotations manuelles. Enfin, nous adaptons

itérativement la PLDA du système de référence avec le corpus cible (traits bleus continus et pointillés).

Le système de SRL nécessite l'apprentissage d'un modèle du monde (UBM), d'un extracteur de *i-vectors* (TV) et d'un modèle PLDA. Leur apprentissage non supervisé nécessite d'extraire des informations sur les locuteurs de façon automatique comme suit. Un UBM à 256 gaussiennes est appris sur des segments de parole détectés grâce à un système parole/non-parole à base de GMMs. L'apprentissage de TV, de rang 200, nécessite des segments contenant un unique locuteur. Ceux-ci sont obtenus grâce à un système de segmentation BIC qui produit, d'après notre expérience, des classes pures. L'apprentissage des paramètres de la normalisation sphérique et du modèle PLDA nécessite des annotations en locuteurs et sessions afin de modéliser les variances inter-et intra-locuteurs. Nous ne considérons, pour apprendre ces paramètres, que les locuteurs apparaissant dans au moins trois émissions pour une durée minimum de 10 secondes par émission. Ces annotations en locuteur sont obtenues grâce à un système non supervisé de SRL similaire à celui décrit à la section 2 mais utilisant une distance cosinus.

4 Contexte expérimental

Les modèles contrastifs pour la SRL ont été appris sur trois corpus distincts, tirés d'environ 200 heures d'émissions audiovisuelles des campagnes d'évaluation REPERE (Galibert & Kahn, 2013), ETAPE (Galibert *et al.*, 2014) et ESTER(Galliano *et al.*, 2009). Ces corpus, où les locuteurs sont identifiés par leur nom et prénom, contiennent plusieurs sessions pour un grand nombre de locuteurs. Les locuteurs apparaissant dans plus d'un épisode sont appelés locuteurs récurrents (r.), par opposition aux locuteurs ponctuels (p.), qui ne sont présents que dans un épisode.

Nous définissons deux corpus $cible$, construits à partir des corpus officiels d'apprentissage et de test de REPERE (Galibert & Kahn, 2013). Le premier, qu'on appellera LCP_{cible}, est la collection de tous les épisodes disponibles de l'émission $LCPInfo$. Le second, qu'on appellera BFM_{cible}, contient tous les épisodes disponibles de l'émission $BFMStory$. Ces deux émissions ont été choisies parce qu'elles comptent un nombre suffisamment important d'épisodes (plus de quarante chacune), et contiennent de nombreux locuteurs récurrents, qui parlent pour plus de 50% du temps de parole total de la collection. Quelques chiffres à propos des deux corpus sont présentés dans le tableau 1. Les deux corpus étant partiellement annotés, seuls les chiffres des locuteurs annotés sont présentés.

Corpus	LCP_{cible}	BFM_{cible}
Nombre d'épisodes	45	42
Durée d'un épisode	25m	60m
Durée de parole annotée	10h08m	19h57m
Locuteurs ponctuels	127	345
Locuteurs récurrents (2 occurrences ou plus)	93	77
Locuteurs récurrents (3 occurrences ou plus)	48	35
Nombre total de locuteurs	220	422
Locuteurs p., part du temps de parole total	20.12%	44,84%
Locuteurs r. (2+ occ.), part du temps de parole total	79,88%	55,16%
Locuteurs r. (3+ occ.), part du temps de parole total	67,06%	45,94%
Durée de parole moyenne d'un locuteur, par épisode	1m08s	1m58s

TABLE 1 – Composition des deux corpus cibles. Seuls les locuteurs annotés sont présentés. La durée de parole annotée correspond à la durée de parole comptant pour l'évaluation.

Le corpus d'entraînement, ou $train$, est utilisé pour des expériences complémentaires. Il contient 344 enregistrements audios issus des corpus d'entraînement et de développement des campagnes citées précédemment, pour un total de 200 heures de parole annotée. Le corpus contient 3888 locuteurs uniques, dont 391 répondent aux critères minimaux choisis pour l'estimation des paramètres de la PLDA : ils apparaissent dans au moins trois enregistrements avec un temps de parole minimal de dix secondes par enregistrement. Par conséquent, ce corpus est bien adapté pour l'apprentissage d'un système *i-vector/PLDA*.

5 Expériences

La métrique utilisée pour mesurer la performance de SRL est le DER (pour Diarization Error Rate, taux d'erreur de SRL). Il a été introduit par le NIST comme la part de temps de parole qui n'est pas attribuée au bon locuteur, en utilisant la meilleure correspondance entre la référence et l'hypothèse. L'outil d'évaluation utilisé (Galibert, 2013) sert à mesurer la performance de la SRL intra-enregistrement et inter-enregistrements. Dans ce dernier cas, on s'attend à ce que le système identifie chaque locuteur récurrent par la même étiquette dans tous les enregistrements de la collection. Nous reportons les résultats intra-et inter-enregistrements, cependant nous commenterons principalement le DER inter-enregistrements. Les résultats intra-enregistrement montrent dans la plupart des cas un très faible impact du calcul du score entre les paires de locuteurs (tableau 2). Pour le calcul du DER, une erreur de 250ms aux frontières des segments est tolérée et la parole superposée est aussi évaluée.

5.1 $Référence$ et $Oracle$

Le système $référence$ est un système *i-vector/PLDA* supervisé, entraîné sur le corpus $train$ et appliqué sur les deux corpus $cible$. Ce système représente la stratégie usuelle lorsqu'on souhaite traiter un nouveau corpus $cible$ non annoté et qu'on dispose déjà d'un corpus d'entraînement annoté. Pour chaque corpus $cible$, un système $oracle$ est également appris de manière supervisée, à partir des annotations de référence des corpus $cible$.

Enfin, un système est appris de façon non supervisée, uniquement à partir d'annotations automatiques de chaque corpus $cible$. Il s'agit d'un système *i-vector utilisant une distance cosinus* qui ne nécessite pas d'information sur les locuteurs récurrents. Ce système est appelé $référence_{cible}$. Les résultats présentés dans le tableau 2 montrent que sans aucune information a priori sur le corpus $cible$, les performances pour le regroupement inter-enregistrements du système $référence_{cible}$ sont bien plus mauvaises que la $référence$, utilisant des données externes pour l'apprentissage (un DER inter-enregistrements de 29,68% ($référence_{cible}$) contre 17,72% ($référence$) et de 27,62% contre 13,22% respectivement). Ceci s'explique par le fait que le système $référence_{cible}$ ne modélise en aucune façon la variabilité inter-locuteurs.

Si l'on s'intéresse à l'expérience $oracle$, l'apprentissage d'une PLDA supervisée sur le corpus $cible$ n'est pas toujours possible : pour le corpus BFM_{cible}, l'apprentissage de la PLDA ne converge pas, probablement dû à un nombre de locuteurs récurrents trop faible (35). En revanche, lorsque le corpus $cible$ contient suffisamment de locuteurs récurrents, les résultats montrent que l'utilisation de l'information a priori donne les meilleurs résultats avec un DER inter-enregistrements de 10,87%.

5.2 Critères limites pour l'apprentissage de la PLDA

Les résultats $oracle$ montrent que pour un corpus $cible$, il pourrait ne pas y avoir assez de locuteurs récurrents pour apprendre la PLDA. La figure 2 montre que pour le corpus LCP_{cible}, un système PLDA efficace peut être entraîné à partir de 37 épisodes, comprenant alors 40 locuteurs récurrents. Chaque locuteur apparaît alors en moyenne dans 7 épisodes différents. Ces résultats montrent qu'un nombre minimal de locuteurs récurrents est nécessaire pour l'autoapprentissage de la PLDA sur un corpus $cible$. Cependant, la définition des critères minimaux pour l'estimation des paramètres de la PLDA n'est pas encore élucidée. Lorsqu'il y a trop peu de données pour l'apprentissage, l'algorithme EM d'estimation des paramètres ne converge pas.

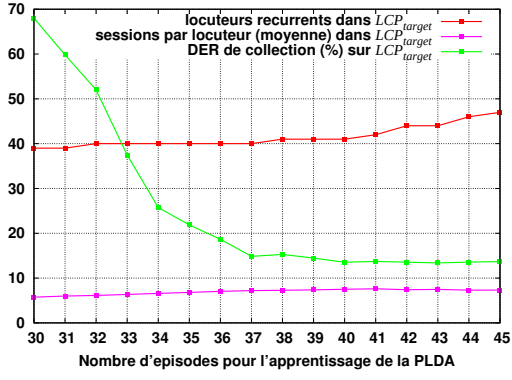


FIGURE 2 – Métriques du regroupement en locuteurs du corpus LCP_{cible}, dont la PLDA est auto apprise avec les annotations. Les valeurs sont fonction du nombre d'épisodes utilisés pour l'apprentissage. Les épisodes ont été sélectionnés dans l'ordre chronologique.

Expérience	UBM + TV appris sur	SNN + PLDA appris sur	LCP_{cible} DER I (%)	LCP_{cible} DER C (%)	BFM_{cible} DER I (%)	BFM_{cible} DER C (%)
$référence$	$train_{sup}$	$train_{sup}$	7,77	17,72	9,92	13,22
$oracle$	$cible_{sup}$	$cible_{sup}$	6,68	10,87	X	X
$référence_{cible}$	$cible_{nonsup}$	sans objet (cosinus)	7,04	29,68	12,46	27,62
$PLDA_{sup}$	$cible_{nonsup}$	$train_{sup}$	7,80	16,58	8,30	12,60
$PLDA_{adapt}$	$cible_{nonsup}$	$train_{sup} + cible_{nonsup}$	7,67	15,60	8,58	11,38

TABLE 2 – DER inter-enregistrement (I) et DER intra-enregistrements (C) des corpus $cible$, pour toutes les expériences.

5.3 Apprentissage non supervisé

Dans la suite de cet article, l'utilisation des annotations fournies avec les deux corpora $cible$ est interdite. Dans (Le Lan *et al.*, 2016), nous avons montré que l'utilisation de sessions issues d'un regroupement BIC pour l'apprentissage non supervisé d'un système *i-vector/PLDA* pouvait fonctionner aussi bien qu'un système supervisé appris sur des données manuellement annotées.

Nous avons donc entraîné un UBM et une matrice TV à partir des sessions BIC. Sur cette base, nous avons appris une PLDA supervisée sur les données du corpus $train$ complémentaire (expérience $PLDA_{sup}$). Les résultats, présentés dans le tableau 2, montrent que l'utilisation d'un UBM et d'une matrice TV appris sur les données $cible_{nonsup}$ à la place des données $train_{sup}$ améliore le DER inter-enregistrements d'environ 1% absolu (16,58% contre 17,72% et 12,60% contre 13,22%).

Dans un second temps, nous avons appris un modèle PLDA à partir des locuteurs issus de la sortie du système de SRL de l'expérience $PLDA_{sup}$. En effet, certaines classes sont étiquetées comme locuteurs récurrents par le système de SRL. L'apprentissage de la PLDA sur ces données n'a pas fonctionné, l'algorithme EM ne convergeant pas. Ceci s'explique probablement par le fait que ces classes ne sont pas suffisamment pures, ne permettant pas d'agréger des statistiques suffisantes à la convergence de l'algorithme EM.

La composition des données pour l'apprentissage supervisé et non supervisé des UBM, TV et PLDA pour les deux corpora est présentée dans le tableau 3. La quantité de données utilisées pour l'apprentissage supervisé est deux fois moins importante que pour l'apprentissage non supervisé. On rappelle que seulement 50% des corpus sont annotés, alors que le système de SRL automatiquement est appliqué sur l'intégralité de l'émission. Cette différence se traduit aussi par un nombre plus important de locuteurs pour l'apprentissage non supervisé.

	LCP_{cible}		BFM_{cible}	
	sup	$nonsup$	sup	$nonsup$
Données utilisées pour l'apprentissage de UBM/TV	9h56m	19h17m	19h50m	39h09m
Durée moyenne d'une session pour l'apprentissage de UBM/TV	1m08s	4m23s	1m58s	4m10s
Nombre de classes-locuteurs pour l'apprentissage de la PLDA	47	130	35	190
Nombre de sessions moyen par classe-locuteur	7,31	5,25	5,45	4,34
Durée moyenne d'une session	1m10s	1m25s	2m50s	2m07

TABLE 3 – Composition des données d'apprentissage de l'UBM, TV et PLDA, pour les deux corpus $cible$.

5.4 Adaptation au domaine

Nous avons constaté que les classes produites par la SRL à l'expérience $PLDA_{sup}$ ne sont pas suffisantes pour apprendre une PLDA dans tous les cas. L'expérience $PLDA_{adapt}$ consiste à utiliser conjointement le corpus supervisé $train_{sup}$ et le corpus cible non supervisé $cible_{nonsup}$. On constate une amélioration de la performance, avec une baisse du DER inter-enregistrements de 16,58% à 15,60% et de 12,60% à 11,38% pour les deux corpus. Les locuteurs récurrents étiquetés après l'expérience $PLDA_{sup}$ permettent donc d'améliorer la qualité de la modélisation de la variabilité inter-locuteurs.

Dans (Shum *et al.*, Proc Odyssey 2014), la question de l'adaptation au domaine est abordée par une combinaison de données annotées externes et de données non annotées acoustiquement proche des données cibles, mais sans utiliser les données cibles elles-mêmes. Les travaux cités consistent à introduire une variable de pondération entre les données internes et externes pour l'estimation des paramètres de la PLDA. Notre approche est plus simple, elle consiste à concaténer deux corpus, la pondération dépendant de la taille relative des corpus. Les résultats de nos travaux sont similaires à ceux de l'article cité, avec une amélioration des performances lorsqu'on utilise l'adaptation, ceci malgré la différence de qualité des données utilisées.

La sortie du système $PLDA_{adapt}$ peut être utilisée pour apprendre un nouveau modèle PLDA. Cependant, nos expériences montrent que lorsqu'on itère après l'expérience $PLDA_{adapt}$, le système ne s'améliore plus. L'idée d'apprentissage non supervisé itératif de la PLDA a été introduite dans (Liu *et al.*, 2014), mais il s'agissait d'appliquer un regroupement en locuteurs sur le corpus d'entraînement lui-même, initialisé par une première itération de SRL *i-vector/cosine*. Dans notre approche, les paramètres de la PLDA sont initialisés sur des données externes annotées, puis estimées à nouveau avec l'ajout des données cibles. Si les travaux antérieurs montraient une amélioration des performances en vérification du locuteur avec un apprentissage itératif, notre approche n'a pas été aussi efficace.

Nos données étant des enregistrements non segmentés et multi locuteurs, les *i-vectors* utilisés pour l'apprentissage itératif pourrait ne pas être suffisamment précis en termes de représentation du locuteur. De meilleurs résultats pourraient être obtenus en modifiant le seuil de regroupement inter-émissions. Un seuil plus strict améliorerait la pureté des classes-locuteurs et pourrait permettre une meilleure modélisation PLDA.

6 Conclusion

Dans cet article, nous avons proposé un système de Segmentation et Regroupement en Locuteurs inter-enregistrements utilisant une stratégie d'autoapprentissage pour le traitement de petites collections d'archives audiovisuelles non annotées. Alors que des travaux antérieurs montraient que l'apprentissage non supervisé d'un système *i-vector/PLDA* sur de telles données était valide, la petite taille des corpus cibles ne permet pas un autoapprentissage performant. L'utilisation de données externes pour un premier apprentissage supervisé reste indispensable.

Nous avons appliqué avec succès une technique d'adaptation au domaine, qui s'était avéré efficace pour la tâche de vérification du locuteur sur des données mono locuteurs, afin d'améliorer le système de SRL $référence$. L'utilisation de l'information sur les locuteurs récurrents apportée par une première itération de SRL sur la collection cible a permis d'améliorer le DER sur les deux corpus.

Les travaux futurs seront consacrés à l'amélioration de la méthode d'apprentissage, en introduisant une pondération dans les paramètres de l'adaptation au domaine entre les données externes et les données cibles. Nous prévoyons également d'approfondir l'aspect itératif de l'apprentissage, la littérature ayant montré que des améliorations étaient possibles.

Références

BARRAS C., ZHU X., MEIGNIER S. & GAUVAIN J. (2006). Multi-stage speaker diarization of broadcast news. *IEEE Transactions on Speech and Audio Processing*, **14**(5), 1505–1512.

BOUSQUET P.-M., LARCHER A., MATROUF D., BONASTRE J.-F. & PLCHOT O. (2012). Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis. In *Speaker Odyssey Workshop*, p. 157–164.

DUPUY G., , MEIGNIER S. & ESTÈVE Y. (2014a). Is incremental cross-show speaker diarization efficient to process large volumes of data ? In *Proceedings of Interspeech*, Singapore.

DUPUY G., MEIGNIER S., DELÉGLISE P. & ESTÈVE Y. (2014b). Recent improvements towards ILP-based clustering for broadcast news speaker diarization. In *Speaker Odyssey Workshop*.

FERRÀS M. & BOURLARD H. (2012). Speaker Diarization and Linking of Large Corpora. In *Proceedings of IEEE Workshop on Spoken Language Technology*, Miami, Florida (USA).

GALIBERT O. (2013). Methodologies for the evaluation of speaker diarization and automatic speech recognition in the presence of overlapping speech. In *INTERSPEECH*, p. 1131–1134.

GALIBERT O. & KAHN J. (2013). The first official repere evaluation. In *Proceedings of Interspeech satellite workshop on Speech, Language and Audio in Multimedia (SLAM)*.

GALIBERT O., LEIXA J., GILLES A., CHOUKRI K. & GRAVIER G. (2014). The ETAPE Speech Processing Evaluation. In *Conference on Language Resources and Evaluation*, Reykyavik, Iceland.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Proceedings of Interspeech*, Brighton, Royaume Uni.

GHAEMMAGHAMI H., DEAN D., VOGT R. & SRIDHARAN S. (2012). Speaker attribution of multiple telephone conversations using a complete-linkage clustering approach. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, p. 4185–4188 : IEEE.

KARAM Z. N. & CAMPBELL W. M. (2013). Graph embedding for speaker recognition. In *Graph Embedding for Pattern Analysis*, p. 229–260. Springer.

KENNY P. (2010). Bayesian speaker verification with heavy tailed priors. In *Speaker Odyssey Workshop*.

LE LAN G., MEIGNIER S., CHARLET D. & DELÉGLISE P. (2016). Speaker diarization with unsupervised training framework. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on* : IEEE.

LIU W., YU Z. & LI M. (2014). An iterative framework for unsupervised learning in the plda based speaker verification. In *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*, p. 78–82 : IEEE.

MEIGNIER S., BONASTRE J.-F. & MAGRIN-CHAGNOLLEAU I. (2002). Speaker utterances tying among speaker segmented audio documents using hierarchical classification : towards speaker indexing of audio databases. In *INTERSPEECH* : Citeseer.

SHUM S. H., CAMPBELL W. M., REYNOLDS D. *et al.* (2013). Large-scale community detection on speaker content graphs. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, p. 7716–7720 : IEEE.

SHUM S. H., REYNOLDS D. A., GARCIA-ROMERO D. & MCCREE A. (Proc. Odyssey 2014). Unsupervised clustering approaches for domain adaptation in speaker recognition systems.

TRAN V.-A., LE V. B., BARRAS C. & LAMEL L. (2011). Comparing multi-stage approaches for cross-show speaker diarization. In *INTERSPEECH*, volume 201, p. 1053–1056.

VAN LEEUWEN D. A. (Proc. Odyssey 2010). Speaker linking in large data sets.

YANG Q., JIN Q. & SCHULTZ T. (2011). Investigation of cross-show speaker diarization. In *INTERSPEECH*, volume 11, p. 2925–2928.

Bilinguismes et compliance phonique

Marie Philippart de Foy[1], Véronique Delvaux[1, 2], Kathy Huet[1], Myriam Piccaluga[1], Rima Rabeh[1] & Bernard Harmegnies[1]

(1) Institut de Recherche en Sciences et Technologies du Langage, Service de Métrologie et Sciences du Langage, Université de Mons, Mons, Belgique

(2) Fonds National de la Recherche Scientifique, Bruxelles, Belgique

Marie.Philippartdefoy@umons.ac.be ; Bernard.Harmegnies@umons.ac.be

RESUME

BILINGUISMES ET COMPLIANCE PHONIQUE

Certains types de bilinguisme pourraient avoir un impact positif sur l'apprentissage phonique et faciliter l'acquisition d'une L3. Certains bilingues pourraient donc présenter une meilleure compliance phonique (aptitude à produire des sons de parole non familiers) que les monolingues. Les données de quatre sujets bilingues ont été recueillies lors d'une tâche de reproduction de voyelles synthétiques précédée d'une phase de production de voyelles en langue maternelle (paradigme développé par Huet et *al.*, 2012). Trois indices ont été calculés et comparés à ceux obtenus par des monolingues francophones lors d'une étude précédente (Delvaux et *al.*, 2014). Les résultats n'ont pas révélé de différence significative entre monolingues et bilingues. Toutefois, le classement des bilingues variait d'un indice à l'autre, suggérant des profils plus diversifiés que chez les monolingues. En conclusion, ces résultats confirment la complexité de la compliance phonique, en particulier chez des locuteurs bilingues, et soulignent l'intérêt d'une approche multi-componentielle ainsi que le besoin d'ajustements ultérieurs de la réflexion théorique sous-jacente.

ABSTRACT

BILINGUALISMS AND PHONETIC COMPLIANCE

Certain types of bilingualism may have a positive impact on sound learning and could facilitate the acquisition of a L3. Some bilinguals could thus demonstrate enhanced phonetic compliance (inherent ability to produce unusual speech sounds) in comparison to monolinguals. Data of four bilingual participants were collected in a reproduction task involving synthesized vowel-like stimuli preceded by typical realisations from both native languages (paradigm developed by Huet et *al.*, 2012). Three indices were computed and compared to those of monolingual participants of a previous study (Delvaux et *al.*, 2014). Results revealed no significant difference between monolinguals and bilinguals. However, the ranking of bilingual participants differed across the three indices, suggesting more diversified profiles than among monolinguals. In conclusion, these results confirm the complexity of phonetic compliance, particularly in bilingual speakers, and emphasize the interest of a multi-componentional approach as well as the need for further refinements of the theoretical underlying reflection.

MOTS-CLES : compliance phonique, bilinguisme, apprentissage d'une L3, aptitude en langues étrangères, capacités individuelles, talent phonétique.

KEYWORDS: phonetic compliance, bilingualism, L3 learning, foreign language aptitude, individual skills, phonetic talent.

1 Introduction

Cette étude porte sur l'impact que pourrait avoir l'expérience bilingue sur la compliance phonique d'un locuteur. La notion de compliance phonique désigne l'aptitude intrinsèque d'un locuteur à percevoir et à produire de manière contrôlée des sons de parole non-natifs, autrement dit des sons de parole inhabituels à sa L1 (Delvaux et *al.*, 2014). Définie comme telle, la compliance phonique est considérée comme faisant partie des capacités requises pour acquérir les compétences phonétiques et phonologiques en langue étrangère. Articulée autour de la notion de « talent phonétique » (Jilka et *al.*, 2007), la compliance phonique implique une conception toutefois différente du talent, envisagé ici en tant qu'habileté pouvant s'entraîner et se développer sur base d'une aptitude naturelle élevée (Gagné, 2003). Ainsi, il ne s'agit pas d'une disposition innée mais d'une compétence acquise et en perpétuelle évolution sous l'action à la fois de facteurs linguistiques, cognitifs et psychologiques spécifiques au locuteur et de son expérience particulière avec les langues étrangères (Delvaux et *al.*, 2014). En outre, le concept de compliance phonique et l'approche métrologique développée pour l'évaluer ont été conçus à partir du système monolingue. Dans cette perspective, il serait donc très intéressant d'observer et de comparer les performances des locuteurs issus d'environnements linguistiques divers, et plus particulièrement, de contextes bilingues

En effet, certaines études, quoiqu'en nombre restreint, ont démontré que les bilingues auraient plus de facilité pour apprendre des langues étrangères que les monolingues (Rudgers et Evans, 2015 ; Sanz, 2000 ; Cenoz, 2003). Cette meilleure capacité d'apprentissage pourrait découler de certains avantages cognitifs généralement associés au bilinguisme (Antoniou et *al.*, 2015 ; Cenoz, 2003). Plus précisément, de nombreuses études ont révélé un meilleur contrôle inhibiteur et une attention sélective renforcée – globalement, des fonctions exécutives plus développées – chez les bilingues (pour une revue, voir Barak et Bialystok, 2011). Un autre avantage cognitif très souvent cité est une conscience métalinguistique accrue chez les bilingues, même si certains auteurs préfèrent plutôt parler d'une plus grande conscience "fonctionnelle" du langage résultant d'une expérience d'apprentissage des langues plus conséquente et plus variée (Rudgers et Evans, 2015). En d'autres termes, les bilingues seraient de meilleurs apprenants de L3 parce qu'ils seraient davantage expérimentés dans ce domaine que les monolingues. D'une manière globale, le lien entre bilinguisme et développement cognitif serait donc suceptible d'engendrer à son tour un impact positif sur les capacités de traitement des sons de parole (Antoniou et *al.*, 2015). Précisons cependant que l'avantage cognitif bilingue et son effet positif sur l'apprentissage d'une L3 dépendraient du type de bilinguisme et du contexte sociolinguistique dans lequel l'expérience bilingue s'inscrit (Sanz, 2000). En effet, si le bilinguisme est un phénomène individuel et social (Sanz, 2000), il est également multidimensionnel et changeant (Hamers et Blanc, 2000). Par conséquent, il n'existe pas une mais plusieurs formes de bilinguismes (pour une des typologies du bilinguisme, voir Romaine, 1989), reflétant des expériences linguistiques diverses et résultant en des niveaux de compétence linguistique variés. Dès lors, nous pouvons supposer que ces différents types de bilinguisme n'engendreront pas les mêmes effets sur la compliance phonique d'un locuteur.

De manière plus spécifique et directement en lien avec le sujet de cette étude, il a été également observé que les locuteurs bilingues pourraient avoir un avantage au niveau de la perception des contrastes phonémiques non-natifs et d'après certains auteurs, cet avantage dépendrait de la difficulté "universelle" et du degré de similarité de ces contrastes avec ceux présents dans les langues natives du locuteur (Antoniou et *al.*, 2015). Dans cette optique,

certains traits phonétiques seraient plus difficiles à apprendre que d'autres et une plus grande similarité phonétique avec les langues natives faciliterait l'apprentissage phonétique de tels sons de parole (Antoniou et *al.*, 2015). Toutefois, deux modèles théoriques, le Speech Learning Model (SLM) de Flege (2007) et le Perceptual Assimilation Model étendu à l'apprentissage d'une L2 (PAM-L2) de Best et Tyler (2007), suggèrent au contraire qu'en cas de similarité avec les sons de la L1, les locuteurs auraient plus de difficulté à atteindre une perception des sons de L2 comparable à celle de locuteurs natifs et ce, par un effet d'assimilation aux catégories phonologiques de la L1. En outre, il existe des résultats plus contradictoires quant à la preuve d'une plus grande flexibilité de traitement de l'information phonétique chez les locuteurs bilingues (Werker, 1986). Beach et *al.* (2001), quant à eux, attirent l'attention sur le lien entre perception et production et soulignent également l'impact des différences individuelles, au-delà des antécédents linguistiques, sur les compétences de perception et de production des locuteurs bilingues.

Pour toutes les raisons que nous venons d'énoncer, nous pouvons supposer que certains types de bilingues pourraient présenter une meilleure compliance phonique que les monolingues ou du moins, que cette compétence ne se développerait pas de la même manière chez des locuteurs bilingues et monolingues. Jusqu'à présent, la compliance phonique a fait l'objet de travaux impliquant des locuteurs monolingues (Huet et *al.*, 2012 ; Delvaux et *al.*, 2014) mais n'a pas encore été étudiée chez des locuteurs bilingues. Par conséquent, nous présentons ici les résultats d'une première étude exploratoire visant à évaluer et à comparer cette compétence chez deux groupes de sujets, des sujets monolingues francophones et des sujets bilingues ayant pour langues natives le français et le néerlandais. Un objectif parallèle sera de déterminer la pertinence et l'adéquation du concept de compliance phonique, tel qu'il a été initialement conçu, vis-à-vis de locuteurs bi ou multilingues.

2 Méthodologie

Le dispositif expérimental de cette étude s'appuie sur celui précédemment développé par Huet et *al.* (2012) et ultérieurement perfectionné par Delvaux et *al.* (2014). Afin de pouvoir évaluer et comparer la compliance phonique chez des adultes bilingues et monolingues, nous avons recueilli un ensemble de données acoustiques auprès de participants bilingues que nous avons ensuite confrontées à celles de participants monolingues d'une étude antérieure (Delvaux et *al.*, 2014).

2.1 Participants

Au total, nous présentons ici les données recueillies auprès de huit participants. Les quatre locuteurs bilingues français-néerlandais sont trois femmes (ultérieurement désignées par B1, B3 et B4) et un homme (ultérieurement désigné par B2) âgés de 21 à 55 ans (moyenne = 30, 7). Trois d'entre eux sont originaires de Belgique et font partie d'une même famille (une mère et deux de ses enfants, respectivement B1, B2 et B3), la quatrième est originaire des Pays-Bas (B4). Tous résident en Belgique, les trois premiers locuteurs depuis leur naissance et la quatrième locutrice depuis 6 ans. Ces participants ont répondu à un questionnaire linguistique afin de déterminer le degré et la nature de leur expérience bilingue. Il ressort de ce questionnaire que les trois locuteurs belges sont des bilingues simultanés et équilibrés et que la locutrice hollandaise est une bilingue tardive qui, de plus, a été régulièrement exposée à l'anglais. Par conséquent, le français et le néerlandais n'ont pas le même statut pour tous nos

locuteurs bilingues. Dans le cas des locuteurs belges, il s'agit de leurs deux langues natives ayant toutes deux le statut de L1. Toutefois, ces bilingues utilisent ces deux langues dans différents contextes et avec différents interlocuteurs ; autrement dit, les deux langues ne sont pas toujours interchangeables. En revanche, le néerlandais est clairement la L1 de la locutrice hollandaise, quant au français, il s'agirait plutôt de sa L2, voire de sa L3. En outre, il nous faut aussi distinguer le néerlandais parlé en Belgique de celui parlé aux Pays-Bas. En effet, il s'agit de deux variétés institutionnalisées du néerlandais standard (Verhoeven, 2005), le néerlandais standard du sud parlé dans les provinces du nord de la Belgique et communément dénommé "Flamand", et le néerlandais standard du nord parlé aux Pays-Bas (Smakman, 2006). Ces deux variétés diffèrent significativement au niveau phonique (Verhoeven, 2005). Qui plus est, en Belgique, il existe également des variétés régionales du néerlandais standard et l'on peut donc parler de contexte diglossique puisque de nombreux locuteurs parlent un régiolecte tout en utilisant le néerlandais dit standard en dehors de la sphère privée et/ou de leur environnement proche (Baetens-Beardsmore, 1980). Les locuteurs d'origine belge et hollandaise ont donc évolué dans des situations sociolinguistiques différentes. Il nous faudra tenir compte de toutes ces informations dans l'analyse de nos résultats. Quant aux sujets monolingues, il s'agit de quatre locuteurs francophones natifs, deux femmes et deux hommes âgés de 31 à 42 ans (moyenne = 34), tous originaires de Belgique, dont la connaissance du néérlandais est, au mieux, élémentaire.

2.2. Recueil des données acoustiques

Le paradigme de collecte de données était constitué de deux parties. Dans la première partie, les sujets étaient amenés à produire 10 réalisations pour chacune des voyelles orales de leur(s) langue(s) parlée(s), à savoir pour le français: [i, e, ɛ , a, y, ø, ə, u, o, ɔ], et, pour le néérlandais: [ɪ, i, iː, ʏ, y, yː, eː, ɛ, ɛː, ə, øː, œː, ɑ, aː, ɔ, ɔː, oː, u, uː] (suivant Gussenhoven, 2009). Pour le néerlandais, chaque voyelle était prononcée à deux reprises, dans un mot et de manière isolée. La deuxième partie était une tâche de reproduction (répétée à 6 reprises) de 94 stimuli, avec pour instruction de "répéter le plus fidèlement possible le son entendu, comme s'il s'agissait d'un son d'une langue étrangère". Plus précisément, les stimuli étaient 94 vocoides synthétiques, conçus avec un synthétiseur de Klatt (1980) et uniformément distribués dans un espace F1*F2*F3 aux fréquences mesurées en mels. Les combinaisons de valeurs de F1/F2/F3 ont été fixées en fonction des propriétés formantiques des voyelles de l'ensemble des langues naturelles et respectent les limites de l'espace acoustico-articulatoire humain (Huet et *al.*, 2012 ; Delvaux et *al.*, 2014). En tout, 664 productions ont donc été recueillies par sujet monolingue et 854 par sujet bilingue, pour un total de 5312 voyelles analysées.

2.3. Traitement des données

Pour chaque son, les valeurs fréquentielles des trois premiers formants ont été automatiquement mesurées au milieu des productions à l'aide du logiciel Praat pour être ensuite manuellement vérifiées par des phonéticiens expérimentés. Les centroïdes des clusters formés par les réalisations des phonèmes en français et en néérlandais ont été obtenus en faisant la moyenne de leurs propriétés formantiques sur les 10 répétitions et ce, pour chaque locuteur. A partir de ces données, nous avons ensuite calculé trois indices permettant de quantifier la notion de compliance phonique. Ces trois indices sont complémentaires dans la mesure où chacun d'eux porte sur une facette spécifique de la compliance et en exprime une

quantification distincte (Huet et *al.,* 2012). L'indice 1 consiste à mesurer le degré de ressemblance entre les stimuli et les productions correspondantes en évaluant les déviations de ces productions par rapport aux cibles. En effet, un sujet est dit compliant s'il parvient à produire des sons similaires aux prototypes qui lui ont été présentés. Ainsi, la distance entre la cible et la réponse peut être calculée à partir de la distance euclidienne qui les sépare dans l'espace acoustique tri-dimensionnel F1*F2*F3. L'indice 1 correspond à la somme de toutes les distances, pour tous les **S** stimuli vocoides et les **P** reproductions de chaque stimulus. Il tend vers 0 lorsque les productions se rapprochent le plus de leurs cibles :

$$\text{Indice 1} = \frac{\sum_{s=1}^{S}\sum_{p=1}^{P}\left[\sum_{i=1}^{I}(Fi_{ps} - Fi_{s})^2\right]^{1/2}}{S * P}$$

La notion de distance constitue toujours la base de l'indice 2 mais celui-ci intègre l'existence d'un système phonologique propre au sujet. De fait, la compliance implique de pouvoir se détacher de ses habitudes phoniques, autrement dit de s'aventurer dans des zones de l'espace vocalique non utilisées dans les réalisations de sa langue native. Le but de cet indice est donc d'incorporer un facteur exprimant l'éloignement de la production par rapport aux clusters vocaliques de la L1 du sujet (en l'occurrence ici, le français, langue commune à tous nos participants), c'est pourquoi on va pondérer les distances entre stimuli et réponses en fonction de leur singularité par rapport aux réalisations habituelles du locuteur. Ainsi, la distance Euclidienne entre les reproductions et leur cible respective va être multipliée par le logarithme de la somme de toutes les distances entre une production et le centroïde du cluster de chaque voyelle **v**. Toutefois, c'est l'inverse de la distance qui sera pris en compte ici, de telle sorte que le résultat de l'indice soit positivement corrélé avec la compliance. Au final, au plus la production ressemble au stimulus et s'éloigne des zones présentes en français, au plus le produit obtenu sera grand et la compliance du sujet sera bonne.

$$\text{Indice 2} = \frac{\sum_{s=1}^{S}\sum_{p=1}^{P}\left\{\prod_{v=1}^{V}\log\left[\sum_{i=1}^{I}(Fi_{ps} - \overline{Fi_{v}})^2\right]^{1/2}\left[\sum_{i=1}^{I}(Fi_{ps} - Fi_{s})^2\right]^{-1/2}\right\}}{S * P}$$

L'indice 3 propose une approche plus statistique : il consiste à analyser la variabilité des variances des distances euclidiennes entre chaque cible et les réponses associées. Si, lors de la tâche de reproduction, les reproductions d'un locuteur varient aléatoirement autour de la cible visée mais que cette variabilité reste constante quelle que soit la cible, ce locuteur est considéré comme compliant. Si, au contraire, le locuteur est fortement influencé par son système phonologique natif, la dispersion de ses réalisations aura tendance à varier en fonction de la cible visée (par effet d'aimantation perceptuelle). Plus concrètement, la variance sera plus ou moins grande selon que la cible est plus ou moins éloignée d'une zone de l'espace vocalique prototypique d'une voyelle de sa L1 (Delvaux et *al*., 2014). L'indice 3 tendra donc vers 0 si le sujet est compliant.

$$\text{Indice 3} = \text{var}_S\left\{\sum_{p=1}^{P}\text{var}_P\left(\left[\sum_{i=1}^{I}(Fi_{ps}-Fi_{s})^2\right]^{1/2}\right)\right\}$$

3 Résultats

La **Table 1** présente les résultats obtenus aux trois indices de compliance phonique (I1, I2 et I3) ainsi que le classement (du plus compliant '1' au moins compliant '4') des locuteurs bilingues et monolingues (issus de l'étude de Delvaux et *al.*, 2014) pour chacun de ces indices.

Locuteur	*I1*		*I2*		*I3*	
	Valeur	Rang	Valeur	Rang	Valeur	Rang
B1	223	4	72	3	4370	3
B2	158	2	64	4	2121	1
B3	184	3	81	2	7111	4
B4	149	1	101	1	4303	2
M1	200	3	62	4	3782	3
M2	216	4	75	3	7457	4
M3	148	2	80	2	2457	2
M4	137	1	87	1	1552	1

Table 1 – Valeur des trois indices pour les 4 sujets bilingues (B1, B2, B3, B4) et monolingues (M1, M2, M3, M4).

La **Figure 1** affiche dans un espace acoustique F1*F2 (mels) les propriétés acoustiques des stimuli et réponses obtenues lors de la tâche de reproduction, ainsi que des centroïdes des voyelles orales du français, pour les quatre locuteurs bilingues. Elle permet d'observer et de comparer la répartition des productions de chaque sujet sur l'ensemble de son triangle vocalique ainsi que d'illustrer les différences de performance capturées par les trois indices.

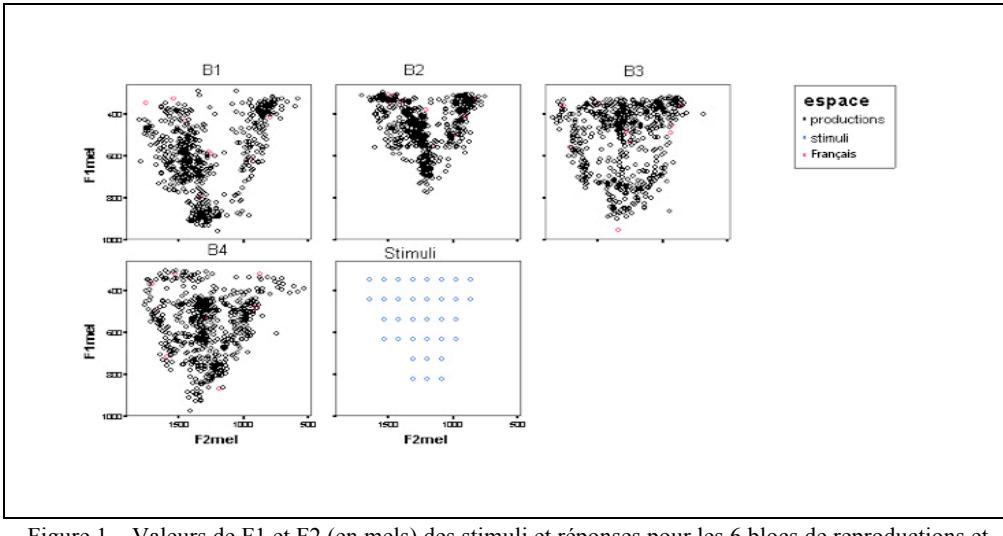


Figure 1 – Valeurs de F1 et F2 (en mels) des stimuli et réponses pour les 6 blocs de reproductions et centroïdes des 10 phonèmes-voyelles orales du français (en rouge) pour les quatre locuteurs bilingues.

Si l'on observe les valeurs des indices des deux groupes de locuteurs, elles n'apparaissent pas particulièrement différentes d'un groupe à l'autre. En effet, les valeurs d'indice 1 et 3 sont globalement plus petites chez les locuteurs monolingues et à l'inverse, l'indice 2 est globalement plus élevé chez les bilingues. Toutefois, ces différences sont minimes et les valeurs d'indice des deux groupes sont en fait très proches. Par ailleurs, nous avons réalisé un test non-paramétrique, le test U de Mann-Witney, afin de comparer les deux groupes de locuteurs sur chaque indice, mais aucune différence significative n'a été mise en évidence. Contrairement à notre hypothèse de départ, les résultats n'ont donc pas révélé de différence significative entre monolingues et bilingues, ce qui signifie que ces derniers ne seraient pas plus compliants. Toutefois, si l'on analyse les résultats en termes d'indices et de rang, on observe des différences notables entre les deux groupes. Le classement des participants bilingues varie en effet davantage d'un indice à l'autre que celui des locuteurs monolingues qui est relativement cohérent sur les trois indices. Plus concrètement chez les locuteurs monolingues, M3 et M4 obtiennent les meilleurs scores et gardent systématiquement le même rang pour les trois indices, M4 étant la plus compliante et M3 venant en deuxième position. M1 et M2 sont donc derniers pour tous les indices et occupent le même rang si ce n'est que M1 se place devant M2 pour les indices 1 et 3 et derrière M2 pour l'indice 2 (pour une description plus complète des sujets monolingues, voir Delvaux *al.*, 2014).

Dans le groupe bilingue[1], on observe en revanche qu'aucun des locuteurs n'obtient la même position dans le classement pour les trois indices. En effet, les résultats de B2 et B3 ne sont absolument pas convergents, B1 et B4 obtiennent quant à eux le même rang pour deux des trois indices. Premièrement, B2 obtient le plus haut rang pour l'indice 3, il est donc le locuteur le plus stable dans ses productions. Toutefois il est dernier pour l'indice 2, ce qui indique qu'il sort très peu de ses habitudes phoniques. Et de fait, si l'on observe les productions de B2 sur la Figure 1, elles apparaissent fortement concentrées dans certaines zones qui sont en rapport avec celles de ses réalisations de voyelles du français. Cette plus grande constance au niveau de la variabilité des variances chez B2 pourrait en quelque sorte être la conséquence de ses performances peu exploratoires. A l'inverse, B3 est dernière pour l'indice 3, elle démontre donc la plus grande variabilité de variance mais sa performance est meilleure pour l'indice 2, pour lequel elle occupe le deuxième rang, indiquant une plus grande prise de risque dans ses productions. Celles-ci sont en effet plus dispersées et mieux réparties sur l'ensemble de son triangle vocalique (Fig.1). B1 apparait en dernière position pour l'indice 1 et en avant-dernière pour les indices 2 et 3, elle est donc globalement la moins compliante des locuteurs bilingues. Ses productions sont relativement regroupées dans certaines zones, certaines d'entre elles coïncidant avec celles des réalisations en L1 (Fig.1) Aussi, il semblerait que l'espace vocalique de B1 comporte des zones totalement non utilisées. Enfin, B4 est quant à elle le sujet le plus compliant, excepté pour l'indice 3 où elle se place juste derrière B2. A l'instar de B3, ses productions sont assez dispersées sur l'espace vocalique tout en présentant des zones de forte concentration, tout comme chez B2, si ce n'est que ces zones ne correspondent pas aux centroïdes des voyelles du français produites par B4 (Fig.1). Ceci indique que B4 est la locutrice qui prend le plus de risques et réalise les productions les plus éloignées de ses habitudes phoniques en français, ce qui occasionne une plus grande variabilité de variance.

[1] Pour rappel, B1, B2 et B3 sont de la même famille et originaires de Belgique alors que B4 est originaire des Pays-Bas ; nous reviendrons sur ce point dans la discussion.

4 Discussion

L'objectif de cette étude était d'observer l'impact du bilinguisme sur la compliance phonique, aptitude à produire des sons de parole non inhabituels au(x) langue(s) native(s) du locuteur. Plus concrètement, nous avions supposé que les locuteurs bilingues pourraient être, dans une certaine mesure, plus compliants que leurs pairs monolingues et ce, grâce à de meilleures capacités de perception et production de sons de parole, possiblement dues à un avantage cognitif plus général. Afin de pouvoir vérifier cette hypothèse, nous avons comparé les performances de locuteurs bilingues et monolingues et plus concrètement, les valeurs de trois indices calculés sur base d'un ensemble de données acoustiques permettant de quantifier différentes facettes de la compliance phonique (d'après Huet et *al.,* 2012 et Delvaux et *al.*, 2014). Nos résultats n'ont pas révélé de différence significative entre les valeurs d'indice des deux groupes, ce qui signifierait que les locuteurs bilingues ne sont pas plus compliants que les monolingues. Toutefois, ces résultats se révèlent être plus complexes qu'il n'y parait au premier abord. En effet, contrairement aux locuteurs monolingues dont le classement est exactement le même sur les trois indices (si ce n'est une inversion de position entre deux sujets pour l'indice 2), le classement des bilingues varie d'un indice à l'autre et ce, chez les quatre locuteurs bilingues. Leurs résultats sont donc moins convergents en termes de rang, ce qui suggère moins de cohérence que chez les monolingues. En d'autres termes, les profils des locuteurs bilingues seraient plus diversifiés et complexes que ceux des monolingues. La Figure 1, où apparaissent les valeurs formantiques de l'ensemble des données acoustiques recueillies pour chaque sujet bilingue, permet d'ailleurs de visualiser ces différences de performances capturées par les trois indices. L'absence de différence significative au niveau des valeurs d'indice des deux groupes ne signifie donc pas que les locuteurs monolingues et bilingues font montre d'une compliance phonique similaire.

Comme nous l'avons déjà précisé, les trois indices ont été initialement construits dans l'objectif de mesurer la compliance phonique chez des locuteurs monolingues. En effet, l'indice 2 implique la prise en compte de la structure du système phonologique natif du locuteur dans l'évaluation de sa performance à la tâche de reproduction ; toutefois, dans le cas d'individus bilingues, il faudrait alors considérer le système phonologique des deux langues parlées. Mais, si l'on intègre l'existence des deux systèmes phonologiques, on augmente le nombre de clusters vocaliques, autrement dit le nombre de zones desquelles un locuteur doit se distancier pour obtenir une bonne valeur d'indice 2. Par conséquent, si nos sujets bilingues ont également produit 10 réalisations des voyelles orales du néerlandais, nous avons choisi de n'incorporer que les données du français afin de ne pas d'emblée les désavantager par rapport aux monolingues et de permettre une meilleure comparaison des deux groupes. Mais, ce faisant, nous avons occulté une partie de l'information et les valeurs d'indice des sujets bilingues ont en quelque sorte été biaisées. C'est d'ailleurs pour cela que, dans l'ensemble, les locuteurs bilingues semblent être légèrement meilleurs pour l'indice 2 et ceci explique probablement aussi le fait que les résultats diffèrent d'un indice à l'autre en terme de rang pour tous les locuteurs bilingues. Par conséquent, il sera nécessaire de perfectionner l'indice 2 ou de développer un indice 2 bis impliquant le choix de dix centroïdes de clusters vocaliques, parmi les clusters vocaliques des différentes langues du locuteur, correspondant aux zones de son espace vocalique les plus utilisées, afin d'adapter notre approche métrologique à des locuteurs bi ou multilingues.

De plus, notre échantillon de locuteurs est également trop restreint pour être à même de faire des généralisations et qui plus est, les locuteurs bilingues n'ont pas tous exactement le même

profil. En effet, trois d'entre eux (B1, B2 et B3) provenant d'une même famille et originaires de Belgique, sont des bilingues simultanés et équilibrés ; la quatrième locutrice (B4) est, quant à elle, originaire des Pays-bas, et présente un bilinguisme de type tardif. Or, comme nous l'avons souligné précédemment, l'avantage que pourraient avoir les bilingues pour l'acquisition d'une L3, et plus spécifiquement pour l'apprentissage phonétique, pourrait ne pas s'étendre à tous les types de bilinguisme (Sanz, 2000) ainsi qu'à toutes les paires de langues (Antoniou et *al.*, 2014). Aussi, nous n'avons pas pris en compte le statut socio-culturel des participants, celui-ci pouvant également avoir une incidence sur les résultats. Nos résultats suggèrent donc que l'hétérogénéité du bilinguisme pourrait avoir un lien avec la plus grande variabilité en termes de compliance phonique chez les locuteurs bilingues par rapport aux monolingues. En effet, les facteurs qui influent sur le fonctionnement bilingue sont multiples. Il en résulte un impact différentiel en fonction du type de bilinguisme et ce, sur les différentes facettes de la compliance phonique ; en d'autres termes, certains bilingues seraient avantagés pour certains aspects de la compliance et d'autres moins, ou pour d'autres aspects. Il se pourrait, par exemple, qu'une expérience bilingue précoce dans un contexte diglossique n'engendrerait pas forcément une plus grande flexibilité puisque les locuteurs auraient appris très tôt à maintenir leurs deux systèmes linguistiques séparés. En outre, B4 a également été régulièrement exposée à l'anglais et le français serait davantage sa L3 que sa L1. Par conséquent, il n'est pas surprenant que cette locutrice s'éloigne le plus aisément des clusters vocaliques du français puisqu'elle est, en quelque sorte, la moins francophone des quatre locuteurs bilingues. Par ailleurs, B4 se démarquait également des autres sujets de par son attitude extrêmement positive vis-à-vis de la tâche, ce qui pourrait également avoir amélioré sa performance.

En conclusion, les résultats de cette étude confirment la multidimensionnalité et la complexité de la compliance phonique, en particulier chez des locuteurs bilingues. En effet, analyser la compliance chez des individus bi ou multilingues nécessite de prendre en compte à la fois les caractéristiques spécifiques au locuteur et à son environnement, son expérience avec les langues étrangères ainsi que la combinaison particulière des langues qu'il parle. De plus, une bonne compliance phonique n'est pas forcément corrélée avec ce qui est habituellement considéré comme un "bon" bilinguisme, c'est-à-dire un bilinguisme précoce et équilibré. La relation entre bilinguismes et compliance phonique est donc complexe et variable, c'est pourquoi ces notions doivent être appréhendées de manière subtile et nuancée. Par conséquent, cette étude souligne l'intérêt d'une approche multi-componentielle dans l'évaluation de la compliance phonique ainsi que le besoin d'ajustements ultérieurs de la réflexion théorique sous-jacente et du paradigme expérimental afin de pouvoir évaluer et comparer cette compétence chez des locuteurs issus de n'importe quel environnement linguistique.

Références

ANTONIOU M., LIANG E., ETTLINGER M., WONG P. (2014). The bilingual advantage in phonetic learning. *Bilingualism: Language and Cognition*, 18(4), 683-695.

BARAC R., BIALYSTOK E. (2011). Cognitive development of bilingual children. *Language Teaching*, 44(1), 36-54.

BEACH E. F., BURNHAM D., KITAMURA C. (2001). Bilingualism and the relationship between perception and production: Greek/English bilinguals and Thai bilabial stops. *International Journal of Bilingualism*, 5(2), 221-235.

BEATENS BEARDSMORE H. B. (1980). Bilingualism in Belgium. *Journal of Multilingual &*

Multicultural Development, 1(2), 145-154.
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*, 13-34.
CENOZ J. (2003). The additive effect of bilingualism on third language acquisition: A review. *International Journal of Bilingualism*, 7(1), 71-87.
DELVAUX V., HUET K., PICCALUGA M., HARMEGNIES B. (2014). Phonetic compliance: a proof-of concept study. *Frontiers in psychology*, 5, 1375.
FLEGE J. E. (2007). Language contact in bilingualism: Phonetic system interactions. *Laboratory phonology*, *9*, 353-382.
GAGNE F. (2003). Transforming gifts into talents : the DMGT as a developmental theory. In N. COLANGELO ET G. A. DAVIS (Eds.), *Handbook of Gifeted Education,* Boston, p. 60-74.
GUSSENHOVEN C. (2009). Vowel duration, syllable quantity and stress in Dutch. *Hanson, Kristin & Inkelas, Sharon* (eds.), 181-198.
HAMERS J. F., BLANC M. (2000). *Bilinguality and bilingualism*. Cambridge : Cambridge University Press.
HUET K., PICCLAUGA M., DELVAUX V., HARMEGNIES B. (2012). Pour une évaluation de la compliance phonique. Actes des *XXIXe Journées d'Étude sur la Parole JEP2012, Grenoble*, 401-408.
JILKA M., ANUFRYK V., BAUMOTTE H., LEWANDOWSKA N., ROTA G., REITERER S. (2008). Assessing individual talent in second language production and perception. Actes de *New Sounds 2007: Proceedings of the Fifth International Symposium on the Acquisition of Second Language Speech. Florianópolis, Federal University of Santa Catarina*, 224-239.
ROMAINE S. (1989), *Bilingualism*, London : Blackwell.
RUTGERS D., EVANS M. (2015). Bilingual education and L3 learning: metalinguistic advantage or not?. *International Journal of Bilingual Education and Bilingualism*, 1-19.
SANZ C. (2000). Bilingual education enhances third language acquisition: Evidence from Catalonia. *Applied psycholinguistics*, 21(1), 23-44.
SMAKMAN D. (2006). Standard Dutch in The Netherlands - A sociolinguistic and phonetic description (Doctoral dissertation, Netherlands Graduate School of Linguistics).
VERHOEVEN J. (2005). Belgian Standard Dutch. *Journal of the International Phonetic Association*, 35(2), 243-247.
WERKER J. F. (1986). The effect of multilingualism on phonetic perceptual flexibility. *Applied Psycholinguistics*, 7(2), 141-155.

De bé à bébé : le transfert d'apprentissage auditori-moteur pour interroger l'unité de production de la parole

Tiphaine Caudrelier[1] Pascal Perrier[1] Jean-Luc Schwartz[1] Christophe Savariaux[1] Amélie Rochet-Capellan[1]
(1) Gipsa-Lab, 11 rue des Mathématiques, 38400 Saint Martin d'Hères, France
Tiphaine.caudrelier@gipsa-lab.fr, Amélie.Rochet-Capellan@gipsa-lab.fr

RESUME

La parole est souvent décrite comme une mise en séquence d'unités associant des représentations linguistiques, sensorielles et motrices. Le lien entre ces représentations se fait-il de manière privilégiée sur une unité spécifique ? Par exemple, est-ce la syllabe ou le mot ? Dans cette étude, nous voulons contraster ces deux hypothèses. Pour cela, nous avons modifié chez des locuteurs du français la production de la syllabe « bé », selon un paradigme d'adaptation auditori-motrice, consistant à perturber le retour auditif. Nous avons étudié comment cette modification se transfère ensuite à la production du mot « bébé ». Les résultats suggèrent un lien entre représentations linguistiques et motrices à plusieurs niveaux, à la fois celui du mot et de la syllabe. Ils montrent également une influence de la position de la syllabe dans le mot sur le transfert, qui soulève de nouvelles questions sur le contrôle sériel de la parole.

ABSTRACT

From sensorimotor experience to speech unit.

Speech is often described as a sequence of units associating linguistic, sensory and motor representations. Are these representations linked at the level of a specific unit, for example, the syllable or the word? In the present study, we contrast these two hypotheses. We modified the production of the syllable "bé" (/be/) in French speakers using an auditory-motor adaptation paradigm that consists in altering the speakers' auditory feedback. We studied how this modification then transfers to the production of the word "bébé" (/bebe/). The results suggest a link between linguistic and motor representations both at the word and the syllable level. They also show an effect of the position of the syllable in the transfer word, which raises new interrogations about serial control of speech.

MOTS-CLES : Perturbations, adaptation auditori-motrice, transfert, unité de parole
KEYWORDS: Perturbations, auditory-motor adaptation, transfer, speech unit

1 Introduction

La parole peut être décrite comme une mise en séquence d'unités linguistiques hiérarchisées entre elles, telles que les phrases, les mots, les syllabes ou les phonèmes. Ces unités seraient également de

nature sensorimotrice, c'est-à-dire associant des représentations sensorielles, notamment auditives, et motrices. Au niveau du système moteur, elles correspondent à des gestes articulatoires, qui se succèdent et parfois se recouvrent. Mais quelle est la nature du lien entre ces représentations linguistiques et motrices ? A quel niveau s'effectue la correspondance entre les deux?

Différentes perspectives existent sur cette question. La syllabe Consonne-Voyelle (CV) a souvent été présentée comme l'unité principale de production de la parole. Par exemple la Théorie « Frame then Content » (Cadre puis Contenu) de MacNeilage (1998) suggère que la syllabe CV trouve ses racines dans les oscillations mandibulaires (notamment présentes dans la mastication), et pourrait ainsi être à l'origine de notre langage articulé. Par ailleurs, en psycholinguistique, Crompton (1982) a suggéré l'existence d'un répertoire mental, appelé syllabaire, qui contiendrait les gestes articulatoires nécessaires à la production de chaque syllabe. Levelt (1999) a développé cette idée dans son modèle de production de la parole où la syllabe serait le niveau auquel le message est décomposé afin d'être traduit, à l'aide du syllabaire, en commandes motrices. Il s'appuie sur l'observation des lapsus phonologiques (inversion d'unités dans une phrase), et sur des effets de fréquence des syllabes sur la vitesse de prononciation des mots.

Néanmoins, l'hypothèse que la syllabe serait l'unité majeure de production de la parole est contestée. En particulier, la grande variabilité observée dans la production d'une syllabe par un locuteur en fonction du contexte de communication demande explication. La théorie de l'exemplaire (Medin et Schaffer, 1978) propose une vision alternative permettant de prendre en compte ces variabilités. Elle a d'abord été développée comme un modèle général de perception et de catégorisation, et repose sur le principe que chaque élément perçu d'une catégorie est mémorisé, plutôt que le seul prototype de la catégorie. Ainsi, un stimulus donné serait comparé à tous les éléments en mémoire pour être catégorisé. Cette théorie a été ensuite étendue à l'identification et la reconnaissance de mots par Goldinger (1998) puis à la production de parole (Bybee 2002). Elle est également au cœur des propositions de Vihman et Croft (2007) sur le développement de la phonologie chez l'enfant, dans une perspective où la syllabe émerge de la production des premiers mots.

Notre objectif est d'interroger la nature de l'unité de production de la parole. Pour évaluer quelle unité assure la correspondance entre une séquence de parole et les gestes articulatoires la produisant, une possibilité est de modifier cette correspondance pour une instance donnée, et de voir comment cette modification s'étend à d'autres instances. Par exemple, si la syllabe est l'unité de production assurant le lien avec les commandes motrices, alors si l'on modifie la production d'une syllabe dans un contexte donné, cette modification devrait se généraliser à la production de cette syllabe dans n'importe quel mot. A l'inverse, si le mot est la plus petite unité représentée, alors il ne devrait pas y avoir de transfert à la même syllabe dans un autre mot. Ce sont ces deux hypothèses que nous allons tester, grâce à un paradigme d'adaptation et de transfert de l'apprentissage sensorimoteur.

L'idée de modifier, à l'aide d'une perturbation, la relation entre les unités perceptives et motrices vient de la recherche sur le contrôle moteur, notamment le contrôle du bras (Mattar et Ostry, 2007). Elle a été adaptée à la production de la parole par Houde et Jordan (1998). Ils ont développé le paradigme d'adaptation auditori-motrice, par analogie avec l'étude de la recalibration visuo-motrice consécutive au port de lunettes prismatiques. Le paradigme est le suivant. Les participants doivent répéter un mot de type Consonne-Voyelle-Consonne (CVC) tandis qu'ils entendent leur voix via un casque. Leur retour auditif est modifié en temps réel. L'altération consiste à décaler le formant f1 (et/ou f2) de manière à transformer la voyelle en une autre (par exemple, « head » devient « had »). Les participants modifient alors leur prononciation, en compensant partiellement la perturbation, se rapprochant ainsi de « hid ». Une phase d'entrainement, constituée de nombreuses répétitions,

aboutit à des recalibrations auditori-motrices, ou apprentissage, qui perdurent après l'arrêt de la perturbation. Cet apprentissage peut être évalué en comparant les formants produits avant et après l'entrainement. C'est ce qu'on appelle le post-effet. Houde et Jordan (2002) ont aussi montré que l'apprentissage réalisé sur une certaine instance se transfère partiellement à la même voyelle dans d'autres mots, voire à des voyelles différentes. Le transfert à la même voyelle dans d'autres contextes serait selon eux le signe de l'existence d'une représentation mentale du phonème.

Ce paradigme d'adaptation et de transfert de l'apprentissage auditori-moteur a été repris dans d'autres études, avec des mots ou pseudo-mots de type CVC. Par exemple, dans Rochet-Capellan et al. (2012), chaque groupe de participants est entrainé sur un mot de la forme Consonne-Voyelle-/n/ différent (par exemple « pan » et « ten »), puis le transfert au mot « pen » est évalué. Cette étude montre qu'un transfert significatif intervient pour la plupart des groupes mais que son amplitude dépend du mot d'entrainement. Une généralisation large mais de faible amplitude a été aussi observée en Mandarin, à partir d'une adaptation sur la triphtongue /iau/ (Cai et al., 2010). Purcell et Munhall (2006) ont estimé l'influence de différents paramètres du protocole expérimental (amplitude de la perturbation, durée de l'entrainement) sur l'adaptation et sur la vitesse de disparition des effets de l'apprentissage après arrêt de la perturbation (ou durée du post-effet). D'autres recherches ont contribué à comprendre comment l'adaptation auditori-motrice est influencée par des facteurs linguistiques, tels que le statut lexical, i.e. mot ou pseudo-mot, (Bourguignon et al., 2014) ou la fréquence des mots (Frank, 2011). Cependant, la plupart de ces études ont été réalisées avec des locuteurs de l'anglais. A notre connaissance, l'utilisation de ce paradigme expérimental pour contraster l'importance du mot et de la syllabe CV en production de parole est nouvelle.

2 Méthode

2.1 Participants, équipement et tâche

36 participants âgés de 18 à 35 ans, de langue maternelle française, ont pris part à l'expérience. Ils n'avaient pas de trouble de l'audition, ni de trouble du langage, et étaient naïfs quant au but de l'expérience. Ils ont reçu une carte cadeau de 15 euros pour leur participation. Les locuteurs étaient assis devant un écran, dans une chambre sourde. Ils portaient un casque équipé d'un microphone (Sennheiser HME 26-II-600). Des mots apparaissaient à l'écran pour une durée d'1.1s. Ils avaient pour consigne de lire les mots à haute voix, de manière naturelle, sans murmurer ni crier.

2.2 Design expérimental et stimuli

L'objectif était de comparer le transfert de l'apprentissage auditori-moteur au niveau du mot, et de la syllabe CV. Pour contraster ces 2 niveaux, nous avons choisi d'utiliser pour mot d'entrainement une syllabe CV, et d'étudier d'une part le post-effet pour le même mot monosyllabique, et d'autre part le transfert à un mot de la forme CVCV répétant cette syllabe. Le mot d'entrainement sélectionné était « bé » pour tous les participants. A l'arrêt de la perturbation auditive, un groupe de sujets a prononcé « bé », et l'autre « bébé ». Notre hypothèse est que si la syllabe est l'unité de production principale, on devrait observer un transfert aux deux voyelles de « bébé » de même amplitude que le post-effet observé sur « bé ». A l'inverse, si le mot est l'unité principale, il ne devrait y avoir aucun transfert à « bébé ». Le choix de la syllabe /be/ devait satisfaire à plusieurs critères. Le choix de la voyelle /e/ était contraint par des aspects techniques liés à la perturbation auditive appliquée. La voyelle choisie

devait nécessairement être située entre deux autres voyelles dans l'espace formantique. La détection des formants étant difficile sur les voyelles arrière /o/ et /ɔ/, le choix s'est restreint aux voyelles avant /e/ et /ɛ/. Par ailleurs, la consonne « b » induit une faible coarticulation. En français, la syllabe « bé » est aussi fréquente en première position d'un mot qu'en milieu ou fin de mot. Enfin, la fréquence du mot d'entrainement et de ses voisins auditifs et articulatoires (« bi » et « baie ») peut influencer l'adaptation auditori-motrice, d'après Frank (2011). Nous avons donc cherché à minimiser la fréquence de ces mots.

2.3 Transformation du retour auditif

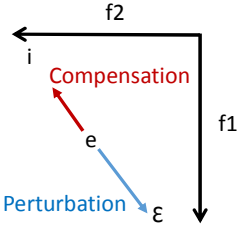


FIGURE 1: La perturbation auditive consiste en un décalage des formants f1 et f2 de manière à transformer le /e/ en /ɛ/ dans le retour auditif des participants. On s'attend à ce que les locuteurs compensent la perturbation en prononçant un /e/ tendant vers le /i/.

Pendant toute l'expérience, les participants parlaient dans un micro et entendaient leur voix dans le casque avec un volume d'environ 70dB, mixée avec un bruit blanc à environ 50dB. Leur voix était enregistrée en parallèle avec un taux d'échantillonnage de 48kHz. Une perturbation en temps réel des formants f1 et f2 était réalisée grâce au logiciel Audapter (Cai et al., 2008). La perturbation a été paramétrée de manière à transformer la voyelle /e/ en /ɛ/. Elle consistait à augmenter f1 de 27% et à diminuer f2 de 10% dans la voyelle /e/. La perturbation créait un retard de 14ms au niveau du retour dans le casque, non perceptible par les participants.

2.4 Procédure expérimentale

Un pré-test permettait d'explorer le triangle vocalique des participants, en leur faisant prononcer des syllabes contenant les voyelles /a/, /ɛ/, /e/, /i/, /u/ et/y/ précédées de la consonne /b/. Le pré-test contenait aussi des paires minimales ayant pour but d'évaluer le contraste /e/-/ɛ/ en production chez les participants (par exemple « épée » vs « épais ») (blocs 1, 2, 3, Fig. 2).

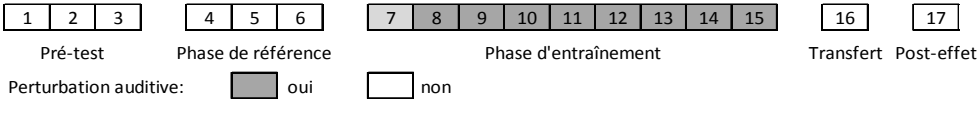


FIGURE 2: L'expérience était constituée d'un pré-test et de 4 phases expérimentales. Une perturbation auditive (en gris) était appliquée pendant la phase d'entrainement.

L'expérience se déroulait par blocs de 20 essais entrecoupés par des pauses. Dans une phase dite de « référence », les participants répétaient 20 fois le mot « bé » (bloc 4, Fig. 2) puis 20 fois le mot « bébé », quel que soit leur groupe (bloc 5). Ensuite, la phase d'entrainement était constituée d'un

bloc de rampe pendant lequel la perturbation était instaurée graduellement jusqu'à son amplitude maximale (bloc 7), suivi de 8 blocs avec la perturbation complète telle que décrite précédemment (blocs 8-15). Tous les participants prononçaient « bé ». A l'issue de l'entrainement, la perturbation était arrêtée, et les participants entendaient à nouveau leur voix non modifiée. Pendant un bloc, dit « de transfert » (16), un groupe continuait à dire « bé » tandis que l'autre groupe prononçait « bébé ». Enfin, dans le bloc de post-effet (17), les deux groupes prononçaient « bé » afin d'évaluer l'adaptation résiduelle après la phase de transfert.

2.5 Analyse des données

Pour chaque voyelle produite, les formants f1 et f2 ont été estimés en prenant une fenêtre de 30 ms au milieu de la partie stable de la voyelle. Pour évaluer l'évolution au cours de l'expérience, les formants ont été exprimés en pourcentage de changement par rapport à leur valeur moyenne pour la même voyelle dans le même mot et pour le même sujet avant la perturbation. Par exemple, le f1 du 2ème /e/ de « bébé » a été rapporté à la valeur moyenne de f1 du 2ème /e/ de « bébé » dans la phase de référence. Comme la perturbation affectait à la fois f1 et f2, la mesure principale choisie pour évaluer le transfert est f2-f1, avec f1 et f2 exprimés en bark (fbark=7*argsinh(fHz/650)).
Si la syllabe est l'unique unité de production de la parole (H1), on ne devrait observer aucun effet du mot de test (« bé » ou « bébé ») ni de la position de la syllabe dans « bébé » sur l'amplitude du changement de f2-f1 observé dans le bloc de transfert. A l'inverse, selon l'hypothèse de spécificité du mot (H2), on devrait observer un post-effet sur « bé » mais pas de transfert à «bébé ». Pour estimer l'adaptation de chaque participant, la valeur de f2-f1 des 40 derniers essais de la phase d'entrainement a été comparée aux valeurs de f2-f1 dans la phase de référence en utilisant un test de Student apparié unilatéral. Le transfert a été estimé en faisant un test de Student sur la valeur du changement sur f2-f1 par rapport à la phase de référence, en contrastant d'une part le groupe contrôle (prononçant «bé ») avec la 1ère voyelle de « bébé » (test de Student non apparié unilatéral) et d'autre part la 1ère avec la 2ème voyelle de « bébé » (test de Student apparié bilatéral).

3 Résultats

3.1 Adaptation et sélection des participants

Nous nous attendions à ce que les participants compensent la perturbation du retour auditif en augmentant f2-f1 dans leurs productions de « bé » (aboutissant à une syllabe plus proche de /bi/). Comme cette étude se focalise sur le transfert, seuls les 27 participants présentant une adaptation significative ont été inclus dans les analyses, soit un groupe de 13 participants (3 femmes) pour le groupe contrôle (transfert 'bé') et un groupe de 14 participants (3 femmes) pour le groupe test (transfert 'bébé').

L'évolution moyenne de f2-f1 au cours de l'expérience est représentée sur la Figure 3. En réponse à la diminution de f2-f1 dans le retour auditif, on observe une augmentation progressive de f2-f1 en production qui se stabilise en fin d'apprentissage. L'adaptation a été évaluée en comparant les 40 derniers essais de la phase d'entraînement avec les 20 essais de la phase de référence. En moyenne, on observe une augmentation de f2-f1 de 7.4% (erreur standard : ±0.3%). Celle-ci ne dépend pas du groupe. Elle correspond à une compensation de la perturbation d'environ 25%. Ainsi, l'adaptation est significative et homogène entre les deux groupes. L'amplitude de la compensation est comparable à celle qui a été observée dans des travaux antérieurs (Rochet-Capellan et al., 2012).

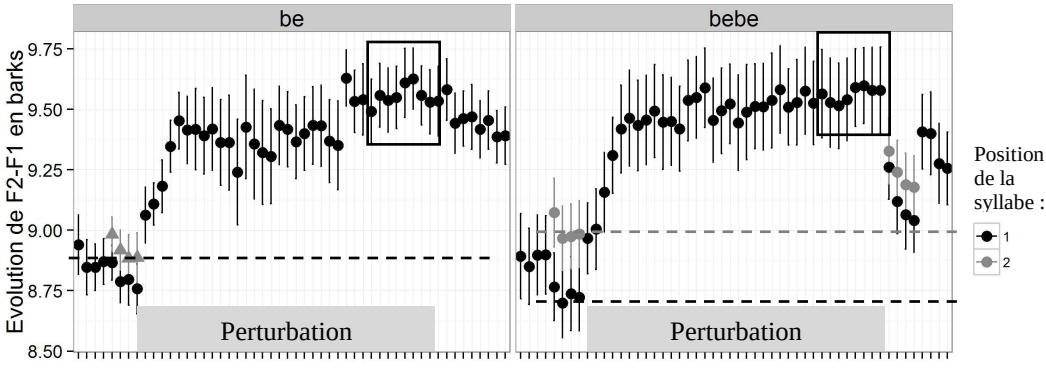


FIGURE 3: Evolution de f2-f1 en barks au cours de l'expérience, pour les sujets adaptés de chaque groupe (à gauche 'bé', à droite 'bébé'). Chaque point représente la moyenne du groupe pour 5 essais avec l'intervalle de confiance. Les lignes pointillées donnent les valeurs de la phase référence pour 'bé' et 'bébé'. Les rectangles montrent les essais considérés pour quantifier l'adaptation.

3.2 Post-effet sur « bé » vs transfert sur « bébé »

Le transfert a été mesuré en comparant les 20 essais qui suivent l'arrêt de la perturbation du retour auditif (bloc 16 de la Fig. 2) avec les valeurs de f2-f1 de la phase de référence, pour chacun des mots et des syllabes. Les résultats sont représentés pour chaque groupe, mot et syllabe, sur la Figure 4. L'amplitude du changement de f2-f1 sur « bé » dans le groupe contrôle est de 7.0% (±1.0%). Pour le groupe ayant prononcé « bébé », le changement par rapport à la phase de référence est respectivement de 4.6% (±0.7%) pour la 1ère syllabe et 2.6%(±0.7%) pour la 2ème syllabe. Un t-test révèle que le post-effet sur « bé » est significativement plus élevé que le transfert à la 1ère syllabe de « bébé » ($t(25) = 1.97$, $p = 0.03$). De plus, le transfert à la 1ère syllabe de « bébé » est significativement plus grand que le transfert à la 2ème syllabe ($t(13) = 3.25$, $p = 0.006$).

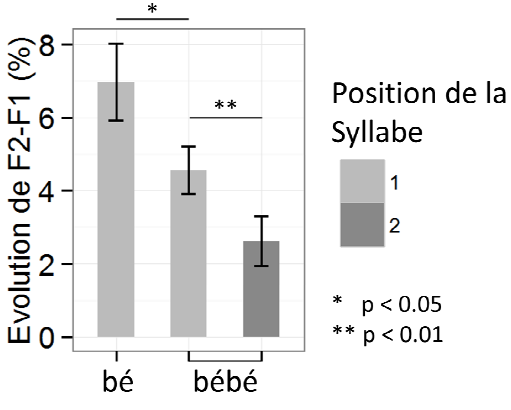


FIGURE 4: f2-f1 moyen dans le bloc de transfert (16), en pourcentage de changement par rapport aux valeurs de référence, par groupe (« bé » ou « bébé ») et par syllabe (1ère ou 2ème position).

Dans le bloc de post-effet (bloc 17), l'amplitude du changement de f2-f1 sur « bé » par rapport à la phase de référence a été mesurée pour les deux groupes. Chez le groupe qui a prononcé « bé » dans le bloc précédent, le post-effet est de 6.1%(±1.0%). Il est de 5.0%(±0.6%) chez le groupe ayant prononcé « bébé ». Il n'y a pas de différence significative entre les deux groupes (p=0.27).

4 Discussion

Avant de discuter des résultats sur le transfert, il est important de préciser un certain nombre d'hypothèses sur la nature de l'adaptation sensorimotrice. D'abord, on peut se demander si l'adaptation est de nature sensorimotrice ou bien uniquement perceptive. Si l'adaptation était uniquement perceptive, alors le locuteur s'habituerait à un /e/ qui ressemble à un /ɛ/ (c'est le sens de la perturbation auditive). On pourrait donc s'attendre à ce que lorsque l'on arrête la perturbation, le participant se mette à prononcer une syllabe proche de /bɛ/, pour rester dans la même catégorie perceptive. Or c'est l'inverse qui se produit. Le locuteur prononce une syllabe proche de /bi/, qui est éloignée d'un point de vue auditif mais identique d'un point de vue moteur à ce qu'il a appris pendant l'entrainement. Il ne s'agit donc pas d'une adaptation des catégories auditives.

D'autre part, on peut également se demander si l'adaptation résulte d'une recalibration automatique chez les participants, ou bien si ces derniers adoptent une stratégie de compensation de la perturbation, en supposant ce que l'expérimentateur attend d'eux. Les participants ont répondu à un questionnaire à la fin de l'expérience. On leur demande notamment s'ils ont perçu quelque chose de bizarre au niveau de leur retour auditif, s'ils ont eu l'impression que leur voix avait été modifiée, et si oui comment. Sur 36 participants, 21 ont déclaré avoir remarqué une modification de leur voix, et seulement 10 participants (soit 28%) ont mentionné une transformation de /e/ vers /ɛ/. Selon 5 d'entre eux, cette transformation était occasionnelle (par exemple au début de certaines séquences). Ainsi la plupart des participants n'avaient aucune idée de ce que l'expérimentateur pouvait attendre d'eux, puisqu'ils n'avaient même pas identifié la perturbation. Par ailleurs une étude a montré que les participants s'adaptent à une perturbation auditive même quand on leur demande de ne pas s'adapter (Munhall et al., 2009).

4.1 Que nous apprend le transfert sur l'unité de production de la parole?

Notre objectif était de contraster les rôles respectifs du mot et de la syllabe en production de parole, en étudiant le transfert de l'apprentissage sensorimoteur. Les résultats représentent un mélange de nos deux hypothèses. Le transfert significatif aux voyelles de « bébé » plaide pour l'existence d'une représentation d'une unité de production de la parole plus petite que le mot, qui pourrait être la syllabe CV. Néanmoins, la différence d'amplitude entre le post-effet sur « bé » et le transfert à « bébé » questionne l'existence d'un syllabaire mental qui représenterait l'unique lien entre une séquence de parole cible et les gestes articulatoires permettant de la produire. Ainsi, le lien se ferait à de multiples niveaux, incluant à la fois le mot et la syllabe, et probablement d'autres unités.

Il est à noter que les sujets du groupe « bébé » ont prononcé la syllabe « bé » deux fois plus que les autres dans le bloc de transfert. On pourrait donc s'attendre à ce que les effets de l'entrainement disparaissent plus rapidement dans ce groupe. Néanmoins cet effet possible de la répétition n'est pas à lui-seul à l'origine de l'effet observé entre le post-effet sur « bé » et le transfert à « bébé ». En effet, le post-effet mesuré dans le bloc suivant est équivalent dans les deux groupes.

4.2 Effet de la position de la syllabe

On observe dans cette étude une influence de la position de la syllabe dans le mot sur le transfert de l'apprentissage sensorimoteur. Cette différence va de nouveau à l'encontre de la conception d'un syllabaire contenant une représentation unique, un prototype, de la syllabe et des gestes articulatoires qui lui sont associés. En effet, le transfert est significativement plus élevé sur la 1ère syllabe que sur la 2ème syllabe. Nous proposons plusieurs explications à cette observation.

D'abord, cet effet de position pourrait être lié à une influence de la prosodie sur le transfert de l'apprentissage sensorimoteur. Une syllabe plus accentuée pourrait être produite de manière plus précise. Nous avons observé chez les participants de cette étude que la première syllabe de bébé est significativement plus accentuée que la 2ème selon deux indices prosodiques : la fréquence fondamentale f0 et l'énergie acoustique sont plus élevées sur la 1ère syllabe que sur la 2ème. A l'inverse, la durée de la 2ème voyelle est plus longue que la 1ère. Ces observations sont valables à la fois dans la phase de phase de référence et dans le bloc de transfert.

L'effet de position observé pourrait dépendre aussi de la perturbation. La phase d'entrainement a été réalisée avec un mot monosyllabique. La syllabe perturbée était donc la première (et l'unique). Dans ce contexte, le locuteur pourrait peut-être apprendre à corriger particulièrement la 1ère syllabe d'un mot ou d'une séquence. Aucune étude utilisant le paradigme d'adaptation auditori-motrice n'a été réalisée à notre connaissance chez l'humain avec des mots de plusieurs syllabes. Cependant, une étude du transfert de l'apprentissage auditori-moteur dans une séquence de syllabes a été réalisée chez les oiseaux. Hoffman et Sober (2014) ont utilisé le paradigme d'adaptation auditori-motrice sur des vocalisations de diamants mandarins. Ceux-ci étaient équipés d'un mini casque leur fournissant un retour auditif. Une perturbation de leur fréquence fondamentale était appliquée sur une syllabe dans une position précise de leur vocalisation. Cette perturbation a induit une adaptation allant dans le sens d'une compensation, comme chez l'humain. Après l'arrêt de la perturbation, un transfert a été observé pour les syllabes identiques à la syllabe sur laquelle la perturbation avait été appliquée, quelle que soit leur position dans la séquence. Le transfert s'étendait partiellement aux syllabes voisines dans la séquence de test. Chez l'humain, l'utilisation de perturbations sélectives en termes de position dans le mot ouvrirait de nouvelles perspectives intéressantes sur l'étude des programmes moteurs nécessaires à la réalisation de séquences de parole.

Remerciements

Ces recherches ont bénéficié du soutien financier 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).

Références

BYBEE, J. (2002). Phonological Evidence for Exemplar Storage of Multiword Sequences. *Studies in Second Language Acquisition*, *24*.

CAI, S., BOUCEK, M., GHOSH, S. S., GUENTHER, F. H., & PERKELL, J. S. (2008). A System for Online Dynamic Perturbation of Formant Trajectories and Results from Perturbations of the Mandarin. *International Seminar on Speech Production 2008*, 65–68.

CAI, S., GHOSH, S. S., GUENTHER, F. H., & PERKELL, J. S. (2010). Adaptive auditory feedback control of the production of formant trajectories in the Mandarin triphthong /iau/ and its pattern of generalization. *The Journal of the Acoustical Society of America*, *128*, 2033–2048.

FRANK, A. F. (2011). Integrating Linguistic, Motor, and Perceptual Information in Language Production. *Dissertation Abstracts International, B: Sciences and Engineering*, *72*, 2454.

GOLDINGER, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

HOFFMANN, L. A., & SOBER, S. J. (2014). Vocal Generalization Depends on Gesture Identity and Sequence. *Journal of Neuroscience*, *34*(16), 5564–5574.

HOUDE, J. (2002). Sensorimotor Adaptation of Speech I : Compensation and Adaptation. *JSLHR*, *45*(April 2002).

HOUDE, J. F., & JORDAN, M. I. (1998). Sensorimotor adaptation in speech production. *Science (New York, N.Y.)*, *279*(1998), 1213–1216.

LEVELT, W. J. M. (1999). Models of word production. *Trends in Cognitive Sciences*, *3*(6), 223–232.

MACNEILAGE, P. F. (1998). The frame/content theory of evolution of speech production. *The Behavioral and Brain Sciences*, *21*(4), 499–511; discussion 511–546.

MATTAR, A. A G., & OSTRY, D. J. (2007). Modifiability of generalization in dynamics learning. *Journal of Neurophysiology*, *98*(6), 3321–3329.

MEDIN, D. L., & SCHAFFER, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*(3), 207.

MUNHALL, K. G., MACDONALD, E. N., BYRNE, S. K., & JOHNSRUDE, I. (2009). Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate. *The Journal of the Acoustical Society of America*, 125(1), 384–390.

PURCELL, D. W., & MUNHALL, K. G. (2006). Adaptive control of vowel formant frequency: evidence from real-time formant manipulation. *The Journal of the Acoustical Society of America*, *120*, 966–977.

ROCHET-CAPELLAN, A., RICHER, L., & OSTRY, D. J. (2012). Nonhomogeneous transfer reveals specificity in speech motor learning. *Journal of Neurophysiology*, *107*, 1711–1717.

VIHMAN, M. M., & CROFT, W. (2007). Phonological development: Toward a “radical” templatic phonology. *Linguistics*, 45, 683–725.

Caractérisation statique et dynamique des voyelles dans des séquences VV.

Julien Millasseau, Olivier Crouzet
Laboratoire de Linguistique de Nantes — LLING — UMR 6310 CNRS / Université de Nantes
Chemin de la Censive du Tertre, 44300 Nantes, France
julien.millasseau@etu.univ-nantes.fr, olivier.crouzet@univ-nantes.fr

RÉSUMÉ

Nous étudions les indices acoustiques liés à la caractérisation statique et / ou dynamique des voyelles du français. Nous avons analysé les caractéristiques formantiques de six réalisations vocaliques ainsi que les transitions formantiques de seize combinaisons V_1V_2 impliquant ces 6 voyelles afin d'évaluer les contributions des indices dynamiques liés aux transitions entre voyelles et des indices statiques de fréquence. Les mesures correspondantes sont issues d'un protocole dans lequel le débit de parole était influencé expérimentalement afin de provoquer d'éventuelles variations de vitesse de transition. Les résultats ne permettent pas de départager ces deux hypothèses mais montrent que les indices dynamiques pourraient être aussi fiables que les mesures statiques. Des pistes d'extension de ce travail sont proposées qui pourraient contribuer de manière plus informative à cette problématique.

ABSTRACT

Static and dynamic characterization of vowels in VV sequences.

The present study aims at evaluating the respective contributions of static and dynamic cues to vowel classification. Formant cues from six french vowels and sixteen V_1V_2 slope transitions were extracted in order to investigate the respective contributions of dynamic and static cues that would be respectively related to transitions or center frequencies. The corresponding data were collected from a dedicated task in which speech rate was influenced experimentally in order to trigger potential variations of rate of change within the transitions. The current results do not favour any of the two potential accounts but show that dynamic cues may be as reliable as static ones. Follow-ups to this protocol are offered that may contribute to this issue more informatively.

MOTS-CLÉS : Dynamique de la parole, Analyse Discriminante Linéaire, Classification des voyelles.

KEYWORDS: Speech Dynamics, Linear Discriminant Analysis, Vowel classification.

1 Introduction

Les études menées en perception et en production de la parole ont montré une grande variabilité articulatoire et acoustique des voyelles (Peterson & Barney, 1952). Ceci est dû à la fois à la diversité des locuteurs, aux variations de débit mais également aux contextes et même aux variations internes à une langue. Certaines études ont soutenu que ces variations étaient mues par la dynamique contextuelle (Lindblom & Studdert-Kennedy, 1967; Strange *et al.*, 1979), et qu'une perspective statique ne pouvait donc suffire à caractériser les voyelles, aussi bien en perception qu'en production. D'autres travaux ont

étendu cette hypothèse aux transitions de séquences CV, VC, CVC (Strange *et al.*, 1983; Strange, 1989; Nearey, 1989; Andruski & Nearey, 1992; Jenkins *et al.*, 1999; Hillenbrand *et al.*, 2001).

Les tâches de perception menées par Strange (Strange *et al.*, 1983; Strange, 1989) et plus tard Jenkins (Jenkins *et al.*, 1999) sur les syllabes dites *silent-center* ont montré qu'une voyelle remplacée par un silence d'une durée équivalente dans une séquence CVC reste aisément identifiable. Selon Strange (Strange *et al.*, 1983; Strange, 1989), ce sont les transitions *on glide* et *off glide* qui permettent, malgré « l'absence » de voyelle, de percevoir ses propriétés linguistiques. Dans cette même optique, Hillenbrand à montré que les éléments les plus pertinents dans une tâche de classification des voyelles sont les valeurs de f_0, la durée ainsi que les valeurs de fréquence de $F1$, $F2$ et $F3$ mesurées dans les portions *onset* (à 25% de la voyelle) et *offset* (à 75%). De ces études ressort la nécessité d'une prise en compte de la composante dynamique dans les tâches de perception et de production de la parole, laquelle semble confirmée par des études récentes mettant en avant la relative stabilité des transitions CV (Hillenbrand *et al.*, 2001) et V_1V_2 (Carré, 2009; Divenyi, 2009) produites par différents locuteurs.

D'après les données de Carré (2009), les pentes maximales des transitions V_1V_2 variraient très peu en fonction des locuteurs et ce même lorsque le débit change. Sur la base des données issues d'une expérience dans laquelle les locuteurs devaient produire une série de séquences V_1V_2 ([aV], dans lesquelles V_1 était systématiquement [a] et V_2 une des 9 voyelles orales du français) à 2 débits différents (« normal »vs. « rapide »), Carré (2009) a comparé les taux de variation des fréquences formantiques mesurées au milieu de chaque V_2 et des pentes maximales mesurées à la transition entre les deux voyelles. Il ressort de la représentation en espace $F1 \sim F2$ que les taux de transition formantique des séquences [aV] seraient soumis à une moins grande variation que les fréquences statiques des $F1$ et $F2$ associés à chaque voyelle. Ces constatations sont fondées sur l'observation des mesures de dispersion des données acoustiques recueillies. Selon Carré (2009), ces indices de pente formantique seraient nécessaires à la caractérisation des voyelles et surtout suffisants pour permettre leur identification malgré les effets de la coarticulation et des variations de débit. Cette hypothèse est soutenue par Divenyi (2009) dans une étude perceptive dans laquelle la vélocité des transitions de séquences V_1V_2 artificielles ressort comme l'une des composantes de la réponse perceptive.

Dans cette étude, nous évaluons la pertinence des indices statiques ($F1$, $F2$, $F3$ et f_0) vs. dynamiques (les pentes « maximales » des transitions) pour la classification vocalique. Cette évaluation recourt à des analyses discriminantes linéaires (LDA pour *Linear Discriminant Analysis*), une méthode statistique de classification basée sur l'utilisation de prédicteurs continus pouvant contribuer à la classification en catégories pré-établies. Ce type d'analyse est principalement exploratoire, au sens où il ne nous indique pas si les indices sont cognitivement pertinents pour un locuteur humain, mais il permet d'analyser la contribution statistique des différents prédicteurs à la séparation des classes impliquées du point de vue de la relation acoustique / linguistique. Nous nous sommes également intéressés au rôle du débit dans la composante dynamique. En effet, selon Carré (2009) le débit n'a pas, ou peu, d'impact sur les pentes des transitions. Or nous pourrions nous attendre à ce que celles-ci s'abaissent à un débit lent et s'accroissent à un débit rapide (O'Shaughnessy, 1986).

2 Méthode

Nous avons donc conçu un protocole de production de la parole nous permettant d'analyser les propriétés acoustiques de différentes séquences de voyelles produites par des locuteurs francophones lorsque le débit varie expérimentalement et d'étudier la contribution de différentes informations

acoustiques à la classification statistique des catégories impliquées. Par manque de place, les effets acoustiques des variations de débit ne seront pas présentés dans cet article.

2.1 Participants

Cinq locuteurs francophones (3 femmes et 2 hommes) âgés de 20 à 25 ans ont participé à cette expérience.

2.2 Matériel

Le corpus de séquences consiste en une série de phrases simples construites autour d'une séquence Voyelle-Voyelle. Nous avons ainsi sélectionné 16 paires V_1V_2 appariées 2 à 2 par leur forme symétrique (/ie/, /ei/, /ia/, /ai/, /iu/, /ui/, /iy/, /yi/, /ea/, /ae/, /eo/, /oe/, /ou/, /uo/, /uy/, /yu/) à partir de 6 voyelles orales du français (/i, e, a, o, u, y/). Les paires ont été construites autour d'une variation de un à trois traits distinctifs (cf. Table 1 ; p. ex. /ie/ : [± haut] ; /iu/ : [± arrière / ± arrondi]). Les seules paires impliquant un changement de 3 traits sont la paire /ia/ et son symétrique /ai/.

	− *arrière* − *arrondi*	− *arrière* + *arrondi*	+ *arrière* − *arrondi*	+ *arrière* + *arrondi*
+ *haut* / − *bas*	*i*	*y*		*u*
− *haut* / − *bas*	*e*			*o*
− *haut* / + *bas*			*a*	

TABLE 1 – Représentation en traits des voyelles utilisées dans l'expérience (inspiré de Durand, 2005).

Ces paires ont servi de base à la sélection de séquences de deux mots par recherche automatisée dans la base de données BRULEX (Content *et al.*, 1990). Le premier mot (toujours un nom commun composé de 2 syllabes) se termine par V_1 alors que le second (toujours un adjectif composé de 3 syllabes) commence par V_2. Par exemple, pour la séquence /eo/ nous avons sélectionné les mots « abbé » et « autonome » (/abeotonom/). Ces suites de deux mots ont ensuite été insérées dans la séquence porteuse « *Regarde [ce|cette|ces] MOT1 MOT2* » (p. ex. « *Regarde cet abbé autonome.* ») afin d'être lues par des locuteurs naïfs. Du point de vue du contexte phonétique, V_1 (resp. V_2) est toujours précédée par (resp. suivie de) une occlusive afin de fournir des marqueurs acoustiques pour la procédure de segmentation. Le voisement et le lieu d'articulation de ces occlusives varient de manière non-contrôlée en fonction des mots. Pour chacune des 16 paires V_1V_2, deux suites nom–adjectif différentes ont été sélectionnées pour aboutir à une liste de 32 séquences Mot1–Mot2.

2.3 Procédure

Les locuteurs devaient lire à haute voix les phrases porteuses présentées aléatoirement sur un écran d'ordinateur par un programme en langage `Python` [1] utilisant la librairie `Pygame` [2]. Chaque locuteur était enregistré en 3 blocs successifs correspondant à 3 débits de parole « cible » (rapide, moyen, lent

1. http ://www.python.org
2. http ://www.pygame.org

— toujours dans cet ordre). Le programme consistait en une barre de progression se développant sous la phrase porteuse affichée, à la manière d'un karaoké. La vitesse de progression de cette barre (elle-même corrélée à l'intervalle inter-stimulus et donc au rythme de succession des phrases) représentait le débit cible vers lequel le locuteur devait tendre, le but étant d'influencer le débit effectif de parole. Lorsque la barre de progression atteignait la fin de la phrase orthographiée, l'écran s'effaçait et une nouvelle phrase apparaîssait après un intervalle de 500ms. Les débits cible « attendus » sont exprimés en ms / syllabe. Les 3 débits cible retenus sont 70ms / syllabe, 140ms / syllabe et 190ms / syllabe. Toutes les phrases sont considérées comme étant composées de 8 syllabes (voyelles non-élidées). Les débits impliqués correspondent donc respectivement aux valeurs temporelles suivantes pour les débits rapide (70ms / syll., durée totale de la progression 560ms, ISI [3] 1060ms), moyen (140ms / syll., progression 1120ms, ISI 1620ms) et lent (190ms / syll., progression 1520ms, ISI 2020ms). Avant chaque bloc, les locuteurs étaient soumis à une phase d'entraînement afin de se familiariser avec l'interface et de s'adapter au débit cible. Pendant la passation, les locuteurs avaient à tout moment la possibilité de mettre le programme en pause. Les enregistrements ont été réalisés dans une pièce silencieuse, à l'aide d'un microphone Røde S1 et d'un enregistreur TASCAM DR-680 (format monophonique digitalisé à 44100Hz et encodé sur 16bits dans un fichier WAV).

Les 1440 réalisations issues des différents enregistrements ont été segmentées et transcrites manuellement à l'aide du logiciel Praat (Boersma & Weening, 2014). Deux tires de segmentation / transcription ont été générées : la première correspond à la séquence V_1V_2 dans son ensemble et est utilisée pour l'analyse « dynamique ». La seconde sert à délimiter les zones stables de chacune des deux voyelles V_1 et V_2 pour l'analyse dite « statique ». Les données acoustiques (f_0 et fréquence des trois premiers formants) associées aux transcriptions indiquées dans le deux tires (classe V_1 ou V_2 ainsi que séquence V_1V_2 correspondante) ont été extraites à l'aide d'un script Praat conçu par le second auteur. Ces données constituent des trajectoires temporelles (position temporelle, f_0, $F1$, $F2$, $F3$) associées à chaque séquence V_1V_2 qui sont ensuite traitées par un programme de manipulation et d'analyse des données dans l'environnement R (R Core Team, 2012). L'extraction de ces trajectoires permet de travailler à la fois sur les valeurs statiques de fréquences formantiques prises à différents instants dans la séquence mais également d'exploiter la composante dynamique en extrayant des valeurs de pente tout au long d'une transition (dérivée en chaque point de la trajectoire).

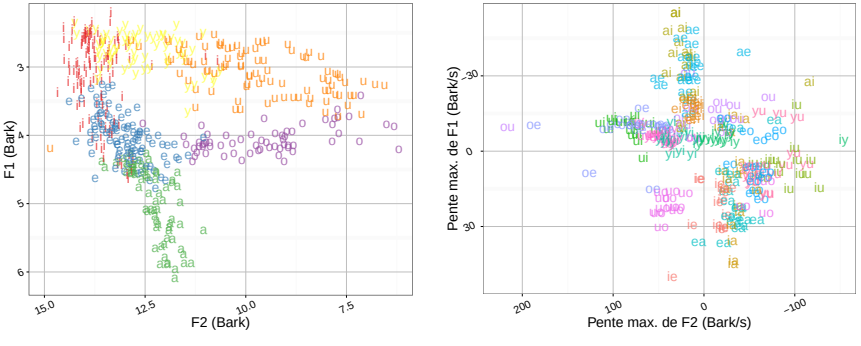
Les données de fréquences ont été transformées en échelle Bark puis normalisées en scores z (variable centrée réduite) pour chaque locuteur (normalisation de Lobanov). Les graphiques présentés dans cet article représentent les données en Bark (ou en Bark/s pour les pentes) afin de faciliter leur lecture. Les analyses statistiques décrites ont toutes été réalisées sur la base des valeurs normalisées, ce qui permet d'améliorer la discriminabilité entre classes dans les Analyses Discriminantes Linéaires. Afin de procéder au traitement des pentes maximales des trajectoires, celles-ci étaient préalablement lissées par une modélisation par courbes splines. La dérivée en chaque point de cette trajectoire était calculée puis la valeur maximale de cette dérivée dans la portion médiane de la séquence V_1V_2 (correspondant à la zone de transition) était extraite. Ceci nous permet d'obtenir une valeur de pente maximale correspondant à la vitesse de transition la plus élevée dans cet intervalle.

3. *Inter-Stimulus Interval* (Intervalle Inter-Stimulus)

3 Résultats

Afin de fournir une première approche des données acoustiques et de leur dispersion préalablement à l'analyse de classification, les fréquences formantiques mesurées au milieu de chaque voyelle et les valeurs de pentes maximales des séquences V_1V_2 produites par le locuteur 2 uniquement sont présentées dans la Fig. 1. La Fig. 1a est classique et permet de visualiser la répartition des réalisations vocaliques pour ce locuteur dans l'espace $F1 \sim F2$. La Fig. 1b représente les pentes maximales mesurées pour chaque séquence réalisée par ce locuteur. On notera que les données globales (les données de l'ensemble des 5 locuteurs) ne changent pas considérablement ces observations subjectives. Les espaces de fréquence des formants sont toujours nettement moins « confus » que ceux des pentes maximales. Les coefficients de variation (écart-type divisé par moyenne) associés aux différentes mesures semblent confirmer cette observation des espaces de variation (Tab. 2).

S'il ressort assez nettement que les données de pente mesurées sont plus difficiles à segmenter en catégories que les données de fréquence des formants, il faut nuancer cette première impression pour deux raisons. Le graphique des fréquences (Fig. 1a) n'est consitué que de 6 classes alors que celui des pentes maximales (Fig. 1b) est composé de 16 catégories. Il est donc prévisible qu'il devrait être plus difficile d'identifier des groupements cohérents « à variation équivalente ». Par ailleurs, si les données servant de base aux arguments de Carré (2009) étaient essentiellement fondées sur les mesures des deux premiers formants, une classification plus visible pourrait apparaître dans un espace multi-dimensionnel impliquant plusieurs mesures acoustiques.



(a) (b)

FIGURE 1 – Représentation (pour le locuteur 2 uniquement) dans un plan $F1 \sim F2$ (a) des valeurs de formants mesurées à 50% de la durée totale de la voyelle (en Bark) ; (b) des valeurs de pentes maximales (en Bark/s).

	Fréquence	*Pente maximale*
F1	*0.16 (0.03)*	*1.71 (1.68)*
F2	*0.12 (0.05)*	*5.43 (20.69)*
F3	*0.04 (0.01)*	*8.75 (18.36)*

TABLE 2 – Moyennes (et écarts-types) des coefficients de variation (écart-type divisé par moyenne) associés à chaque classe (V pour les fréquences, VV pour les pentes).

3.1 Analyses Discriminantes Linéaires

Afin d'approfondir notre compréhension de cette question, et d'explorer plus en détails la piste d'une discriminabilité des classes fondée sur les mesures de vitesse de transition qui pourrait reposer sur un espace multi-dimensionnel, nous avons donc conduit une série d'analyses discriminantes linéaires. Ces analyses pourraient nous permettre de faire ressortir une organisation des données acoustiques qui peut ne pas apparaître dans des espaces en 2 dimensions ou dans des comparaisons de moyennes / médianes et / ou de coefficients de variation.

Les mesures acoustiques de production ont été analysées à travers deux séries d'analyses discriminantes linéaires (LDA). Dans la première série (données statiques), les mesures de $F1$, $F2$, $F3$ et f_0 prélevées au milieu de chaque voyelle (V_1 et V_2) sont utilisées comme prédicteurs des 6 classes vocaliques impliquées. Dans la seconde série d'analyses, les mesures de pente maximale des transitions de $F1$, $F2$, $F3$ et f_0 entre 2 voyelles d'une séquence V_1V_2 sont utilisées comme prédicteurs des 16 catégories de séquences V_1V_2. Notre objectif est d'examiner dans quelle mesure les performances de classification reposant sur des données de pente maximale entre deux segments peuvent être plus « efficaces » que celles qui reposent sur les mesures de fréquence prises au milieu d'un segment.

3.1.1 Données statiques

Le tableau 3 présente les pourcentages de classification correcte des différentes classes vocaliques en fonction des trois groupes de prédicteurs utilisés. Cette analyse nous permet de voir que le pourcentage global de classification correcte varie peu avec l'ajout de prédicteurs. Il faut rentrer en détail dans les catégories vocaliques pour apercevoir un impact, comme celui de $F3$ sur le pourcentage de classification de /y/ pour atteindre un taux global de performance correcte de l'ordre de 70%.

	i	e	a	o	u	y	**global**
$F1 + F2$	78.0	68.2	73.3	66.2	76.5	22.6	**67.2**
$F1 + F2 + F3$	71.9	67.2	74.6	66.2	73.3	60.9	**69.5**
$F1 + F2 + F3 + f_0$	71.9	69.1	73.9	66.2	72.9	60.9	**69.8**

TABLE 3 – Pourcentages de classification correcte associés à chaque voyelle pour les groupes de prédicteurs $F1 + F2$, $F1 + F2 + F3$, $F1 + F2 + F3 + f_0$ (mesure prise au milieu de la voyelle).

.

Afin de pouvoir comparer des taux de performance issus de « tâches » dans lesquelles le nombre de catégories diffère, nous avons appliqué la formule utilisée par Schwartz *et al.* (2004). Cette formule de correction de la proportion de réponses correctes en fonction du nombre de catégories (Éq. 1) permet de représenter la taille relative de la différence entre proportion brute et proportion théorique en fonction du nombre de catégories disponibles.

$$\frac{p - p_{ref}}{1 - p_{ref}} \times 100 \quad (1)$$

Nous appellerons cette taille relative le « score corrigé ». Dans l'équation 1, p représente la proportion de réponses correctes mesurée et p_{ref} la proportion théorique de réponses au hasard ($\frac{1}{6}$ pour 6 catégories par exemple). Les scores corrigés pour le nombre de catégories sont présentés dans le tableau 4. Les données sont évidemment directement comparables au données du tableau 3. Elles seront principalement utiles pour la comparaison avec les données « dynamiques ».

	i	e	a	o	u	y	**global**
$F1+F2$	58.0	48.2	53.3	46.2	56.5	2.6	**44.1**
$F1+F2+F3$	51.9	47.2	53.6	46.2	53.3	40.9	**48.9**
$F1+F2+F3+f_0$	51.9	49.1	53.9	46.2	52.9	40.9	**49.2**

TABLE 4 – Mesures corrigées (Schwartz *et al.*, 2004) des pourcentages présentés dans la table 3.

3.1.2 Données dynamiques

Le but de cette section est d'évaluer la contribution des indices « dynamiques »de pente maximale présents dans les signaux de parole produits par ces locuteurs. Dans cette série d'analyses discriminantes, nous avons cherché à modéliser la classification en 16 classes V_1V_2 en prenant comme prédicteurs les pentes maximales extraites des transitions entre V_1 et V_2. Les mêmes mesures de $F1$, $F2$, $F3$ et f_0 ont servi de prédicteurs mais dans cette partie, ce sont les pentes maximales issues des trajectoires temporelles de ces indices qui ont été utilisées : les pentes maximales correspondent à la vitesse de changement la plus élevée au cours d'une transition. Les données de performance brutes et corrigées sont présentées respectivement dans les tableaux 5 et 6.

	/ie/	/ei/	/ia/	/ai/	/iu/	/ui/	/iy/	/yi/	
$F1+F2$	16.0	76.6	5.1	64.4	72.3	46.0	23.1	47.5	
$F1+F2+F3$	26.7	76.6	20.3	74.7	63.9	55.2	60.0	66.2	
$F1+F2+F3+f_0$	45.3	76.6	45.6	73.6	60.2	62.1	61.5	66.2	
	/ea/	/ae/	/eo/	/oe/	/ou/	/uo/	/uy/	/yu/	**global**
$F1+F2$	67.5	31.0	48.7	23.8	5.6	55.8	24.3	38.5	**41.0**
$F1+F2+F3$	70.1	47.6	46.1	44.0	16.7	48.1	51.4	50.0	**51.4**
$F1+F2+F3+f_0$	72.7	48.8	43.4	40.5	12.5	45.5	52.7	41.0	**53.3**

TABLE 5 – Pourcentages de classification correcte associés à chaque séquence VV pour les groupes de prédicteurs $F1+F2$, $F1+F2+F3$, $F1+F2+F3+f_0$ (pentes maximales de la transition).

.

Le pourcentage brut de classification correcte semble refléter des performances assez nettement inférieures à la classification atteinte à partir des mesures de fréquence statiques. Par ailleurs, on observe de manière similaire à ce qu'on observait sur les données statiques, la présence de certaines classes particulièrement mal catégorisées (/ie/, /ia/, /iy/) pour lesquelles l'ajout de $F3$ améliore parfois sensiblement la classification. La séquence /ou/ ne dépasse cependant jamais les 20% de classification correcte. Dans l'ensemble, le taux maximal de performance correcte brute atteint environ 53%, ce qui est nettement moins élevé que ce qu'on obtenait avec les mesures statiques (environ 70%).

Si l'on se penche sur les scores corrigés par contre, les différences entre les deux ensembles de données s'atténuent fortement. On parvenait à un score corrigé maximal de 49.2% sur la base de mesures de fréquence associées aux 6 catégories vocaliques alors qu'on atteint 46.4% sur la base des pentes maximales associées aux 16 classes de séquences V_1V_2. Ces deux valeurs semblent tout à fait comparables.

	/ie/	/ei/	/ia/	/ai/	/iu/	/ui/	/iy/	/yi/	
$F1+F2$	9.3	70.0	-1.6	57.7	65.6	39.3	16.4	40.8	
$F1+F2+F3$	20.0	70.0	13.6	68.0	57.2	48.5	53.3	59.6	
$F1+F2+F3+f_0$	38.7	70.0	38.9	66.9	53.6	55.4	54.9	59.6	
	/ea/	/ae/	/eo/	/oe/	/ou/	/uo/	/uy/	/yu/	**global**
$F1+F2$	60.9	24.3	42.0	17.1	-1.1	49.2	17.7	31.8	**33.7**
$F1+F2+F3$	63.5	41.0	39.4	37.4	10.0	41.4	44.7	43.3	**44.4**
$F1+F2+F3+f_0$	66.1	42.1	36.8	33.8	5.8	38.8	46.0	34.4	**46.4**

TABLE 6 – Mesures corrigées (Schwartz *et al.*, 2004) des pourcentages présentés dans la table 5.

4 Discussion

Des mesures de variation acoustique et des performances brutes issues des LDA , il ne ressort aucune tendance permettant d'affirmer la moins grande variabilité des pentes transitionnelles par rapport aux fréquences statiques. Au contraire même, les mesures de pente maximale semblent donner lieu à une plus grande variation univariée (cf. les mesures de coefficients de variation associées à $F1$, $F2$ et $F3$ et à leurs pentes maximales) ainsi qu'à des performances de classification correcte nettement moins élevées que celles obtenues à partir des mesures statiques. Cependant nous avons noté que le nombre de catégories possibles entre les LDA « statiques » et « dynamiques » diffère (6 voyelles $\rightarrow$ 16 paires V_1V_2). De cette augmentation du nombre de catégories résulte une diminution du pourcentage théorique de réponse au hasard, aussi le pourcentage théorique de réponse obtenu dans les tableaux 3 et 5, doit être observé par rapport au taux de réponse au hasard, ce qui est proposé à partir des scores corrigés dans les tableaux 4 et 6. Ces résultats ne nous permettent pas de conclure en faveur d'une hypothèse ou d'une autre même si l'écart de performance est très légèrement favorable à l'hypothèse « statique ».

Le prolongement de ces analyses se fera par l'ajout de $F4$ comme prédicteur. Ces mesures sont probablement cruciales pour la classification associée aux voyelles arrondies notamment et pourraient modifier considérablement les résultats des LDA. Nous étudierons aussi en détails les matrices de confusion obtenues afin d'analyser plus précisément la structure des réponses de classification qui pourrait être masquée par les mesures de performance globale. Les méthodes de *resampling* (*bootstrap / permutation*) permettront d'évaluer les intervalles de confiance des mesures de performance observées. Il est par ailleurs délicat d'interpréter une comparaison de résultats de classification impliquant deux ensembles différents de prédicteurs. Nous prévoyons une adaptation de cette procédure qui permettrait d'intégrer dans la même analyse l'ensemble des prédicteurs, ce qui rendrait possible le classement des prédicteurs entre eux.

Ces résultats devront aussi être confrontés aux données perceptives de Divenyi (2009) qui semblent confirmer la prise en compte de la vélocité des transitions dans la perception de séquences de voyelles synthétiques par les locuteurs. Enfin, nous explorerons le traitement de la composante temporelle dans les modélisations par systèmes dynamiques. En effet, dans les méthodes FDA (Functional Data Analysis) une abstraction du temps est nécessaire (Lancia & Tiede, 2012), cela implique donc la suppression de la vitesse or, selon l'hypothèse de Carré (2009), la vitesse serait l'élément essentiel car le plus stable, ce qui semble entrer en contradiction avec la notion d'abstraction temporelle.

Références

ANDRUSKI J. E. & NEAREY T. M. (1992). On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables. *The Journal of the Acoustical Society of America*, **91**, 390–410.

BOERSMA P. & WEENING D. (2014). Praat : Doing phonetics by computer. Computer program. Version 5.4.

CARRÉ R. (2009). Signal dynamics in the production and perception of vowels. In F. PELLEGRINO, E. MARSICO, I. CHITORAN & C. COUPÉ, Eds., *Approaches to Phonological Complexity*, p. 59–81. Berlin – New-York : Mouton de Gruyter.

CONTENT A., MOUSTY P. & RADEAU M. (1990). Brulex. Une base de données lexicales informatisée pour le français écrit et parlé. *L'Année Psychologique*, **90**(4), 551–566.

DIVENYI P. (2009). Perception of complete and incomplete formant transitions in vowels. *The Journal of the Acoustical Society of America*, **126**(3), 1427–1439.

DURAND J. (2005). Les primitives phonologiques : des traits distinctifs aux éléments. In N. NGUYEN, S. WAUQUIER-GRAVELINES & J. DURAND, Eds., *Phonologie et Phonétique : Forme et Substance*, chapitre 3, p. 63–93. Paris : Hermès.

HILLENBRAND J. M., CLARK M. J. & NEAREY T. M. (2001). Effects of consonant environment on vowel formant patterns. *The Journal of the Acoustical Society of America*, **109**(2), 748–763.

JENKINS J. J., STRANGE W. & TRENT S. A. (1999). Context-independent dynamic information for the perception of coarticulated vowels. *The Journal of the Acoustical Society of America*, **106**(1), 438–448.

LANCIA L. & TIEDE M. (2012). A survey of methods for the analysis of the temporal evolution of speech articulator trajectories. In S. FUCHS, M. WEIRICH, D. PAPE & P. PERRIER, Eds., *Speech Planning and Dynamics*, p. 239–271. Frankfurt-am-Main, Germany : Peter Lang.

LINDBLOM B. & STUDDERT-KENNEDY M. (1967). On the role of formant transitions in vowel recognition. *The Journal of the Acoustical society of America*, **42**(4), 830–843.

NEAREY T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, **85**(5), 2088–2113.

O'SHAUGHNESSY D. (1986). The effects of speaking rate on formant transitions in French synthesis-by-rule. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86.*, volume 11, p. 2027–2030.

PETERSON G. & BARNEY H. (1952). Control methods used in a study of vowels. *The Journal of the Acoustical Society of America*, **24**(2), 175–184.

R CORE TEAM (2012). *R : A Language and Environment for Statistical Computing*. Vienna, Austria : R Foundation for Statistical Computing. ISBN 3-900051-07-0.

SCHWARTZ J. L., BERTHOMMIER F. & SAVARIAUX C. (2004). Seeing to hear better : evidence for early audio-visual interactions in speech identification. *Cognition*, **93**, B69–B78.

STRANGE W. (1989). Evolving theories of vowel perception. *The Journal of the Acoustical Society of America*, **85**(5), 2081–2087.

STRANGE W., EDMAN T. & JENKINS J. (1979). Acoustic and phonological factors in vowel perception. *Journal of Experimental Psychology : Human Perception and Performance*, **5**, 643–656.

STRANGE W., JENKINS J. & JOHNSON T. (1983). Dynamic specification of coarticulated vowels. *The Journal of the Acoustical Society of America*, **74**(3), 695–705.

Cartopho : un site web de cartographie de variantes de prononciation en français

Philippe Boula de Mareüil[1], Jean-Philippe Goldman[2], Albert Rilliard[1], Yves Scherrer[2], Frédéric Vernier[1]

(1) LIMSI, CNRS & Université Paris-Saclay, Orsay, France

(2) CUI, Université de Genève, Genève, Suisse

{mareuil;rilliard;vernier}@limsi.fr

{Jean-Philippe.Goldman;Yves.Scherrer}@unige.ch

RESUME

Le présent travail se propose de renouveler les traditionnels atlas dialectologiques pour cartographier les variantes de prononciation en français, à travers un site internet. La toile est utilisée non seulement pour collecter des données, mais encore pour disséminer les résultats auprès des chercheurs et du grand public. La méthodologie utilisée, à base de crowdsourcing (ou « production participative »), nous a permis de recueillir des informations auprès de 2500 francophones d'Europe (France, Belgique, Suisse). Une plateforme dynamique à l'interface conviviale a ensuite été développée pour cartographier la prononciation de 70 mots dans les différentes régions des pays concernés (des mots notamment à voyelle moyenne ou dont la consonne finale peut être prononcée ou non). Les options de visualisation par département/canton/province ou par région, combinant plusieurs traits de prononciation et ensembles de mots, sous forme de pastilles colorées, de hachures, etc. sont présentées dans cet article. On peut ainsi observer immédiatement un /E/ plus fermé (ainsi qu'un /O/ plus ouvert) dans le Nord-Pas-de-Calais et le sud de la France, pour des mots comme *parfait* ou *rose,* un /Œ/ plus fermé en Suisse pour un mot comme *gueule*, par exemple.

ABSTRACT

Cartopho: a website for mapping pronunciation variants in French

This work intends to renew traditional dialectological atlases to map pronunciation variants in French, via a website. The internet is used not only so as to collect data but also to disseminate results to researchers and the general public. The crowdsourcing-based methodology has enabled us to gather information from over 2500 French speakers all over Europe (France, Belgium, Switzerland). A dynamic platform with a user-friendly interface was then designed to map the pronunciation of 70 words in the different regions of the countries under investigation (words including mid vowels or whose final consonants can be pronounced or not, especially). The display options by department/canton/province or region, combining several pronunciation features and sets of words, as coloured pellets, hatches, etc., are presented in this article. One can readily observe a more close-mid /E/ (and a more open-mid /O/) in Nord-Pas-de-Calais and the South of France, for words like *parfait* 'perfect' or *rose* 'pink', a more close-mid /Œ/ in Switzerland for a word like *gueule* 'mouth', for instance.

MOTS-CLES : géolinguistique, dialectologie, cartographie, visualisation, variantes de prononciation

KEYWORDS: geolinguistics, dialectology, mapping, visualisation, pronunciation variants

1 Introduction

La géolinguistique étudie la variation des phénomènes linguistiques dans l'espace, qu'elle peut discrétiser avec un degré de granularité plus ou moins fin. Depuis le début du XX^{e} siècle, époque à laquelle les atlas dialectologiques ont connu un certain intérêt, le paysage linguistique de la France et de l'Europe a considérablement changé. C'est l'espace actuel que nous avons tenté d'appréhender via une expérience à grande échelle, portant sur les variantes de prononciation en français parlé en Europe. Un travail de collecte d'informations a été mené, débouchant sur une importante quantité de données qu'il est difficile de représenter visuellement. En reprenant le principe des atlas linguistiques sur papier et en tirant bénéfice de l'outil informatique, nous avons développé un site web qui permet de cartographier les résultats de manière dynamique. Le site mis au point, cartopho, sera présenté dans ce qui suit.

L'expérience qui a été mise sur pied était centrée sur le timbre des voyelles moyennes (ex. *épée~épais, jeûne~jeune, beauté~botté*) et quelques mots emblématiques, dont en particulier la consonne finale peut être prononcée ou non, selon les régions (ex. *vingt, moins*). Une liste de 70 mots a été établie et lue par un phonéticien, avec à chaque fois deux prononciations possibles. Il était demandé aux participants de préciser laquelle des deux était la plus proche de leur prononciation la plus courante. Des informations complémentaires (notamment sur l'origine et l'ancrage géographique des sujets) ont de plus été demandées. Au total, 2506 sujets, dans tous les départements français, les cantons suisses romands et les provinces belges francophones, ont pris part à cette expérience.

La section suivante introduit la motivation de notre travail, quelques études antérieures et le contexte. La section 3 présente le questionnaire qui a servi à cartographier les variantes de prononciation (pour le français parlé principalement en Europe), la tâche des sujets et les participants. La section 4 décrit l'interface de visualisation, son développement et ses fonctionnalités. Les principaux résultats sont rapportés en section 5. Enfin, la section 6 conclut et ouvre de nouvelles perspectives.

2 Motivation et contexte

La motivation de ce travail est de renouveler les traditionnels atlas dialectologiques pour cartographier la variation phonétique diatopique (i.e. dans l'espace) en français. Il y a plus d'un siècle, l'Atlas Linguistique de la France (ALF) de Gilliéron et Edmont (1902–1910) a permis de produire près de 2000 cartes représentant des données dialectales — les dialectes étant encore assez vivants à l'époque (Goebl, 2002). Plus récemment, le projet Phonologie du Français Contemporain (PFC) a entrepris de recueillir des enregistrements en français dans au moins 36 points d'enquête mais n'a pas véritablement fourni de cartes (Durand *et al*., 2009). Une méthodologie à base de crowdsourcing (c'est-à-dire à grande échelle), aujourd'hui dans l'air du temps (Eskenazi *et al*., 2013), doublée d'outils informatisés de visualisation cartographique permet de combler ce fossé.

Nous nous sommes inspirés d'études antérieures, tel le Harvard Dialect Survey (HDS) développé pour l'anglais américain (Vaux et Golder, 2003) et l'Atlas der deutschen Alltagssprache (AdA) développé pour l'allemand (Elspaß, 2007). Nous avons cherché à représenter sous forme de cartes la variation régionale en français, notamment en ce qui concerne le timbre des voyelles moyennes, dans des mots comme ceux qui sont consignés dans la table 1, où selon la région la prononciation

peut être semi-ouverte ou semi-fermée (Walter, 1976 ; Carton *et al*., 1983). D'autres sources de variation diatopique ont pu être identifiées, comme les voyelles nasales ou le /A/ postérieur, mais nous ont paru plus difficiles à manipuler. Nous nous sommes concentrés sur les voyelles /e/~/ɛ/, /ø/~/œ/ et /o/~/ɔ/, en partie soumises à une loi de position qui est inégalement respectée selon les régions (Walker, 2001 ; Dufour *et al*., 2007 ; Armstrong & Pooley, 2010 ; Boula de Mareüil *et al*., 2013). Cette contrainte phonologique prescrit qu'une voyelle moyenne tend à s'ouvrir en syllabe fermée et, inversement, à se fermer en syllabe ouverte. Elle souffre cependant de nombreuses exceptions, comme nous l'illustrerons à partir du questionnaire en ligne que nous allons à présent décrire.

Syllabe	/E/	/Œ/	/O/
Fermée (CVC(C))	2 (ex. *père*)	10 (ex. *chanteuse*)	26 (ex. *grosse*)
Ouverte (CV)	21 (ex. *lait*)	0	5 (ex. *sot*)

TABLE 1 : Exemples de mots avec l'archiphonème /E/, /Œ/ ou /O/ en syllabe ouverte ou fermée (i.e. se terminant par une voyelle ou une consonne), et nombre d'items à l'intérieur de chaque catégorie dans la liste utilisée pour cette étude.

3 Questionnaire, tâche des sujets et participants

Après des informations sur les participants, concernant le lieu où ils ont grandi, ont passé la plus grande partie de leur vie et vivent actuellement, concernant également leur âge, sexe, niveau d'études, langue maternelle (français oui/non), etc., l'expérience proprement dite consistait à présenter (dans un ordre aléatoire différent pour chaque sujet) une liste de 70 mots, préalablement élaborée et lue par un phonéticien (sans les *e* muets), avec à chaque fois deux prononciations possibles. La table 1 présente le nombre d'items dans chaque catégorie, pour les voyelles moyennes, en syllabe terminée par une voyelle (V) ou une consonne (C). À ces 64 items ont été ajoutés 6 mots emblématiques, dont en particulier la consonne finale peut être prononcée ou non, selon les régions (ex. *vingt*, *moins*) ou qui présentent un [w] à l'initiale (comme en Belgique, à la place d'un /ɥ/ ou d'un /v/ plus canoniques). L'ensemble des mots est inventorié dans la table 2. Il comprend :

- des mots relativement fréquents (dans le corpus PFC), pour lesquels une certaine variation régionale peut être observée ;
- des mots plus rares, (quasiment) absents du corpus PFC, présentant éventuellement quelque hésitation dans les dictionnaires de prononciation pour le français.

/E/	quai près progrès serai serais mais jamais effet piquet inquiet jeunets après dès lait parfait prêt aspect père règle épais épée résonner raisonner
/Œ/	veulent seule aveugle gueule chanteuse meule feutre neutre jeune jeûne
/O/	sauf autre gauche gaufre jaune rose chose synchrone neurone hexagone zone cyclone atome albatros fosse génome grosse arôme bagnole pot sot aurore rhinocéros économe beauté botté vote tome bestiole boussole social
autres	vingt moins encens huit wagon soit (comme dans soit ceci soit cela)

TABLE 2 : Mots utilisés dans l'enquête.

La plateforme Labguistic a été utilisée (Ménétrey et Schwab, 2014). Pour chaque mot, une forme orthographique était fournie aux sujets : pour le mot *gaufre* (*cf*. figure 1), le participant entendait

par exemple une paire de formes comme [gofʁ gɔfʁ] et devait indiquer laquelle des deux formes était la plus proche de sa prononciation la plus courante. Un troisième bouton était proposé, en plus des prononciations 1 et 2 : « je n'entends pas de différence ». Mais nos informateurs ont cliqué sur ce bouton dans moins de 2 % des cas, ce qui nous rassure quant à leur surdité phonologique. À la fin du test, également, des commentaires étaient demandés : les participants étaient invités à répondre à quelques questions sur la difficulté de la tâche.

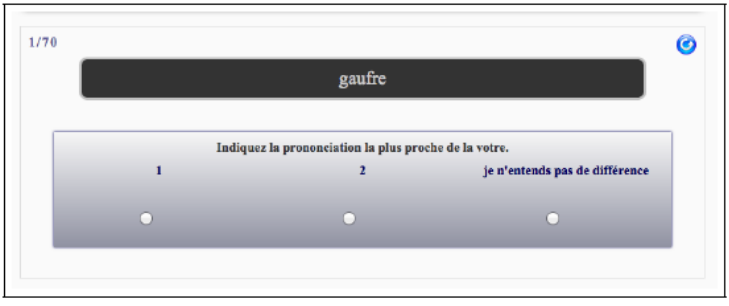


FIGURE 1 : Extrait de l'interface de l'expérience pour le mot *gaufre*.

En quelques mois, nous avons eu plus d'un millier de participants, dont la distribution en termes de tranches d'âge est la suivante : 12 % de moins de 20 ans, 40 % entre 20 et 30 ans, 15 % entre 30 et 40 ans, 11 % entre 40 et 50 ans, 12 % entre 50 et 60 ans, 10 % de plus de 60 ans (Scherrer *et al.*, 2015). Leur distribution géographique, à partir des départements où ils avaient vécu le plus longtemps, est représentée dans la figure 2. Sans surprise, sont surreprésentés des pôles urbains et universitaires en Île-de-France, Rhône-Alpes, Provence-Alpes-Côte d'Azur, ainsi qu'en Belgique et en Suisse.

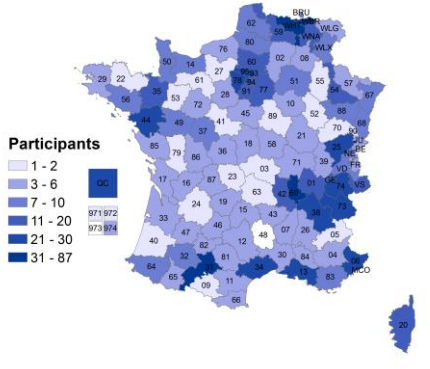


FIGURE 2 : Distribution des départements où les sujets avaient passé la plupart de leur vie.

4 Visualisation : développement et interface utilisateur

Cet ensemble important de données est difficile à visualiser, car il présente de nombreuses dimensions de variation, au-delà de la dimension géographique : les sujets peuvent avoir des histoires de vie plus ou moins complexes (avoir grandi puis longtemps vécu puis déménagé en

différents lieux) et présentent ainsi des caractéristiques sociologiques différentes, qu'il faut pouvoir prendre en compte. Les variables testées durant cette expérience sont multiples et on peut s'attendre à des résultats différents en fonction du caractère plus ou moins emblématique des mots (*rose* ou *vote*) et/ou de la paire de voyelles testée (/O/ ou /E/, par exemple). La question de la représentation des données est donc cruciale. Il faut par ailleurs pouvoir faire évoluer la granularité de la visualisation, si l'on souhaite observer et décrire des phénomènes d'ampleur et de représentativité variables (distinguer par exemple un phénomène spécifique à la Franche-Comté d'un phénomène séparant l'espace francophone en deux). Il faut aussi pouvoir afficher plusieurs phénomène à la fois, afin d'être à même d'observer d'éventuelles différences ou covariations. Toutes ces raisons nous ont poussés à créer une interface web dynamique permettant de représenter une ou plusieurs de ces informations, à différentes échelles d'agrégation géographique et selon diverses caractéristiques sociologiques des informateurs.

Le site cartopho a été développé <https://cartopho.limsi.fr>, disponible en français et en anglais, permettant d'afficher les résultats :

– en fonction du lieu où les informateurs ont grandi, principalement vécu ou résident actuellement ;
– en termes de départements français, provinces belges et cantons suisses ou en termes de régions françaises (avant la récente fusion), Belgique francophone et Suisse romande ;
– avec éventuellement des informations complémentaires comme le nombre de sujets sondés (et le nom du département ou de la région, en cliquant sur la zone correspondante pour faire apparaître une infobulle), un zoom sur la Belgique, la Suisse, etc.

La figure 3 présente un extrait de l'interface utilisateur, dont le menu (à gauche, la carte étant affichée à droite) est divisé en deux parties : options générales en haut, fonctionnalités liées aux variables phonétiques/phonologiques que l'on veut visualiser en bas. L'interface, interactive et conviviale, a vocation à être utilisable par tous (professionnels comme non-spécialistes), permettre une interprétation intuitive des données, à travers des cartes générales ou spécifiques à un mot particulier, suivant les besoins de l'utilisateur. Dans ce but, un panel d'options et un mode d'emploi sont mis à disposition de l'usager ; les cartes et les légendes associées sont mises à jour dynamiquement.

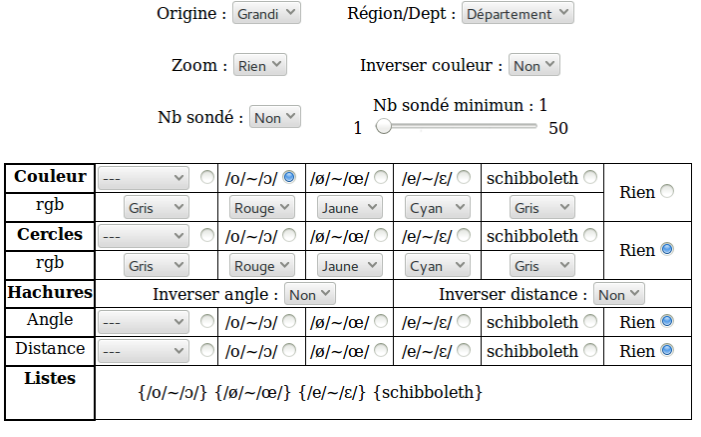


FIGURE 3 : Extrait de l'interface utilisateur du site cartopho (menu par défaut, avec les options générales en haut et des fonctionnalités plus spécifiques en bas).

Des codes couleurs par défaut ont été associés aux résultats, que l'utilisateur peut inverser ou modifier à sa guise, de façon ergonomique : par défaut, une variance de rouge pour la voyelle /O/, de jaune pour la voyelle /Œ/, de cyan pour la voyelle /E/ et de gris pour les autres cas (schibboleths dont notamment le 't' ou le 's' final peut être prononcé). Il est de plus possible de combiner les informations pour plusieurs items, cocher/décocher des mots à l'intérieur de chaque catégorie (mots avec /E/, /Œ/, /O/, etc.), avec des pastilles de couleur ou des hachures (plus ou moins resserrées ou orientées différemment, selon les prononciations)…

Dans le menu (*cf.* figure 3), les lignes « Couleur » et « Cercles » permettent d'afficher les cartes pour un mot ou un ensemble de mot à l'intérieur d'une catégorie (/O/, /Œ/, /E/ ou autre) à l'aide de couleurs de fond ou de pastilles de couleur, respectivement. Chacune des options « Couleur » et « Cercles » est accompagnée d'une ligne « rgb », offrant la possibilité à l'utilisateur de modifier la couleur d'affichage. L'option « Hachures » est quant à elle accompagnée de deux lignes, « Angle » et « Distance », permettant d'afficher des cartes rendant visibles les différences de prononciation à travers l'orientation ou l'espacement des hachures, respectivement. Deux champs « Inverser angle » et « Inverser distance » agissent de la même façon que le champ plus général « Inverser couleur ». La ligne « Listes », enfin, permet de dérouler l'ensemble des mots à l'intérieur de chaque catégorie (/O/, /Œ/, /E/ ou schibboleth) et d'en sélectionner un ou plusieurs : la sélection sera représentée sur la carte sous forme de « Couleur », « Cercles » ou « Hachures », selon les choix de l'utilisateur. Par défaut, plus le timbre d'une voyelle est fermé, plus les couleurs sont foncées, plus les hachures tendent à être orientées horizontalement plutôt que verticalement (option « Angle ») et sont rapprochées (option « Distance »).

Le site a été programmé selon une architecture client-serveur sous différentes plateformes, en PHP et en JavaScript avec le canvas HTML5 comme bibliothèque graphique. Les coordonnées des contours des départements et des régions, les fonctionnalités d'affichage de la carte et de la légende sont encodées dans différents fichiers .js. La base de données résultant du questionnaire (au format CSV et sécurisée), est parsée automatiquement pour retourner l'agrégation nécessaire à la visualisation demandée. Les tableaux de couleurs, calculées selon le système HSL (pour *Hue Saturation Lightness* « Teinte Saturation Luminosité ») ont été créés à partir des nuances proposées par ColorBrewer <http://colorbrewer2.org/>. La figure 4 fournit quelques exemples de cartes générées à partir des départements où les informateurs avaient vécu le plus longtemps.

5 Illustration de quelques résultats

Les cartes de la figure 4 reflètent diverses possibilités d'affichage du site cartopho : avec le nombre de sujets sondés par départements et une infobulle (fig. 4A), avec un zoom sur les cantons suisses (fig. 4D), couplant couleurs de fond et hachures orientées de différentes façons (fig. 4E), combinant couleurs de fond, pastilles de couleur et hachures plus ou moins espacées (fig. 4G). Pour des mots comme *père* (fig. 4A), la carte obtenue est très homogène. Il en va tout autrement pour des mots comme *parfait* ou *jamais* (fig. 4B et 4G) où un [e] fermé prévaut dans le tiers sud de la France. Pour l'archiphonème /Œ/, les choses sont très différentes selon que l'on considère des mots comme *chanteuse* ou *neutre* — où la voyelle est plutôt ouverte dans le sud de la France (à l'exception de la Corse) ainsi que dans la région Nord-Pas-de-Calais (fig. 4C et 4G sous forme de hachures d'autant plus rapprochées que le timbre est ouvert) — ou un mot comme *gueule* — où une prononciation semi-ouverte semble être attestée à travers toute la France, alors qu'on observe une poche dans certains cantons suisse, avec la prononciation semi-fermée [gøl] (fig. 4D).

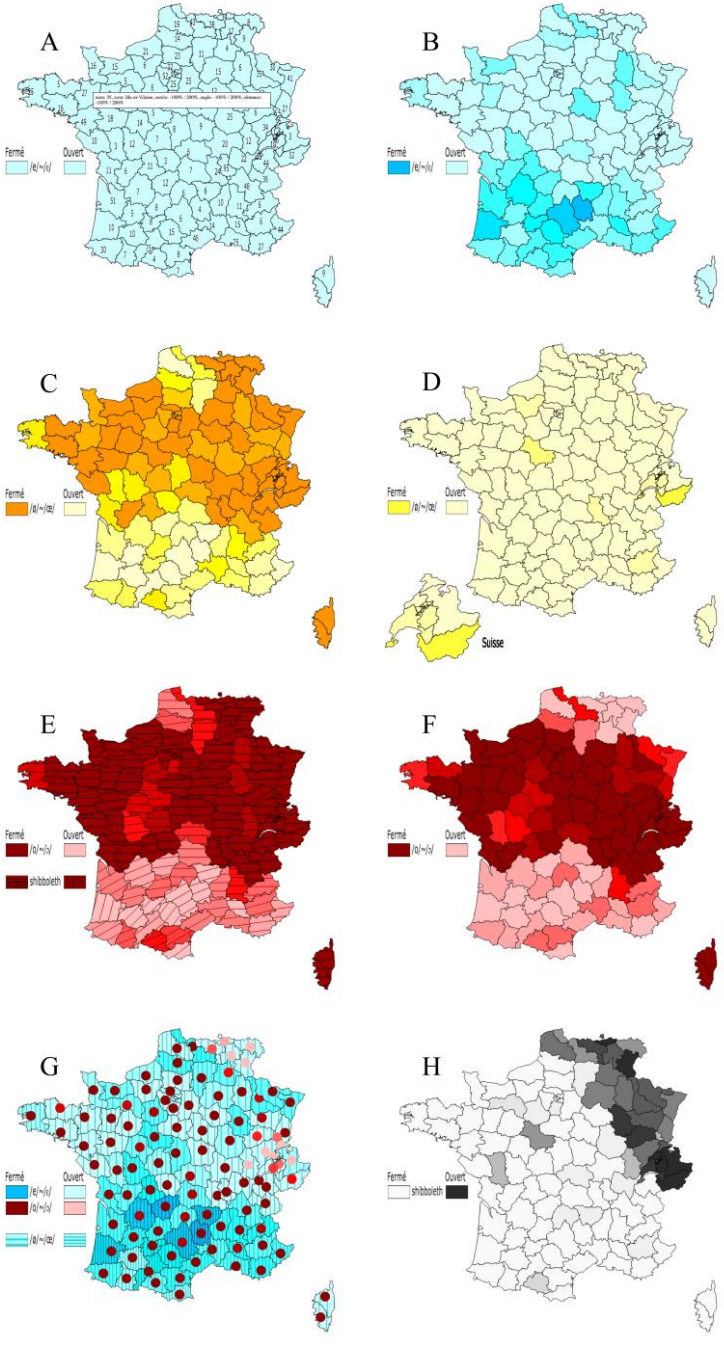


FIGURE 4 : Cartes affichées par le site cartopho pour les mots (A) *père,* (B) *parfait,* (C) *chanteuse,* (D) *gueule,* (E) *rose* + *moins* (hachures), (F) *grosse,* (G) *pot* (pastilles) + *jamais* + *neutre* (hachures), (H) *vingt.*

Pour l'archiphonème /O/, le patron fourni par l'emblématique *rose* (fig. 4E) est assez semblable à celui de *chanteuse* ; de même dans *grosse* (fig. 4F), à la différence près qu'en Belgique la prononciation tend également à être semi-ouverte — comme au masculin, dans l'adjectif *gros*. En syllabe ouverte, la prononciation semi-fermée est majoritaire à travers toute la France (fig. 4G : pastilles rouges). Fait exception toutefois la Franche-Comté, où la prononciation rejoint celle du Jura suisse. Quant à la Belgique, ce que l'on peut noter pour un mot comme *pot* (formant une paire minimale avec *peau*) est conforme à ce que nous venons de dire concernant *gros*, avec une prononciation semi-ouverte [pɔ]. Le test incluait enfin des schibboleths comme *moins* et *vingt* (fig. 4E sous forme de hachures et 4H sous forme de niveaux de gris). On peut constater que dans tout l'arc est de la France (mais également en Suisse et en Belgique) le 't' final de *vingt* tend à être prononcé, alors que c'est dans le Sud-Ouest que le 's' final du mot *moins* se fait entendre.

6 Conclusion et perspectives

En conclusion, on peut aujourd'hui toucher une large audience, à travers des sites grand public ou les réseaux sociaux. Ce travail a montré la faisabilité de grandes enquêtes en ligne, pour étudier et cartographier des variantes régionales de prononciation en français. Le nombre de participants semble suffisant, sauf peut-être dans les régions les moins peuplées, autour de l'Auvergne. De plus, les retours fournis par les sujets en fin de test ont suggéré que la tâche était assez facile. Cette enquête a permis de confirmer des phénomènes connus (comme le fait que la loi de position est mieux respectée dans le sud de la France) et de mettre au jour des phénomènes moins connus (comme des prononciations semi-ouvertes pour des mots tels que *rose* et *grosse* en Nord-Pas-de-Calais et en Belgique). Dans l'imaginaire commun des francophones, de telles prononciations avec [ɔ] sont plutôt associées au Midi. Le site cartopho que nous avons présenté dans cet article permet de visualiser immédiatement ce qu'il en est chez les sujets parlants de différentes variétés de français — si ce n'est à partir de données linguistiques de terrain, du moins sur une base perceptive/déclarative, à partir de jugements sur les prononciations proposées.

Les perspectives qui s'offrent à nous consistent à analyser plus en détail les résultats, à les ventiler pour démêler l'influence de l'âge, du sexe, voire du niveau d'études et de la mobilité des sujets — ce qui peut également être cartographié. Des techniques de classification sont par ailleurs envisageables, pour faire ressortir premièrement la différence Nord/Sud et deuxièmement la distinction entre d'un côté la Suisse, la Belgique et la Franche-Comté, de l'autre (le reste de) la France.

Pour tracer plus précisément des isoglosses, le besoin va se faire sentir de récolter davantage de données dans les zones de transition entre domaines d'oïl et d'oc. Malheureusement, celles-ci correspondent souvent aux régions les moins peuplées de la France, des régions rurales qui se montrent parfois rétives au maillage du territoire — ceci constitue une limite de la méthodologie à base de crowdsourcing. Il s'agira enfin de confronter les prononciations déclarées et les usages réels. Des comparaisons avec les analyses acoustiques menées sur la base du corpus PFC sont en cours et de nouvelles expériences où l'on enregistrerait les sujets sont envisagées.

Remerciements

Nous remercions chaleureusement Nicolas Posada, qui a réalisé le site web cartopho ainsi que les nombreux sujets qui ont pris part à cette expérience à grande échelle.

Références

ARMSTRONG N. et POOLEY T. (2010). *Social and linguistic change in European French*. Basingstoke : Palgrave Macmillan.

BOULA DE MAREÜIL P., WOEHRLING C., ADDA-DECKER M. (2013). Contribution of automatic speech processing to the study of Northern/Southern French. *Language Sciences* 39, 75‑82.

CARTON F., ROSSI M., AUTESSERRE D., LEON P. (1983), *Les accents des Français*, Paris : Hachette.

DUFOUR S., NGUYEN, N., FRAUENFELDER U. H. (2007). The perception of phonemic contrasts in a nonnative dialect. *Journal of the Acoustical Society of America Express Letters* 121, 131‑136.

DURAND J., LAKS B., LYCHE C. (2009). *Phonologie, variation et accents du français*. Paris : Hermès.

ELSPAß S. (2007). Variation and Change in Colloquial (Standard) German ‑ The *Atlas zur deutschen Alltagssprache* (AdA) Project. *In* Fandrych C., Salverda R. (éditeurs), *Standard,Variation und Sprachwandel in germanischen Sprachen/Standard, Variation and Language Change in Germanic Languages*. Tübingen : Narr, 201–216.

ESKENAZI M., LEVOW G.-A., MENG H., PARENT G., SUENDERMANN D. (2013). *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*. Chichester : Wiley.

GILLIERON J. et EDMONT E. (1902‑1910). *Atlas linguistique de la France*. Paris : Champion.

GOEBL H. (2002). Analyse dialectométrique des structures de profondeur de l'ALF. *Revue de linguistique romane* 66(261–262), 5‑63.

MÉNÉTREY P. et SCHWAB S. (2014). Labguistic: a web platform to design and run speech perception experiments. *In* Congosto Martín Y., Montero Curiel M. L., Salvador Plans A. (éditeurs), *Fonética experimental, educación superior e investigación*. Arco/Libros, Madrid, 543–556.

SCHERRER Y., BOULA DE MAREÜIL P., GOLDMAN J.-P. (2015). Crowdsourced mapping of pronunciation variants in European French. *18th International Congress of Phonetic Sciences*, Glasgow, 1‑5.

VAUX B. et GOLDER S. (2003). The Harvard Dialect Survey. Téléchargé de : http://www4.uwm.edu/FLL/linguistics/dialect/.

WALKER D. C. (2001). *French sound structure*. Calgary : University of Calgary Press.

WALTER H. (1976). *La dynamique des phonèmes dans le lexique français contemporain.* Paris : France-Expansion.

Comparaison de listes d'erreurs de transcription automatique de la parole : quelle complémentarité entre différentes métriques ?

Olivier Galibert[1] Juliette Kahn[1] Sophie Rosset[2]
(1) LNE, F-78190 Trappes, France
(2) LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
olivier.galibert@lne.fr, juliette.kahn@lne.fr, rosset@limsi.fr

RÉSUMÉ

Le travail que nous présentons ici s'inscrit dans le domaine de l'évaluation des systèmes de reconnaissance automatique de la parole en vue de leur utilisation dans une tâche aval, ici la reconnaissance des entités nommées. Plus largement, la question que nous nous posons est "que peut apporter une métrique d'évaluation en dehors d'un score ?". Nous nous intéressons particulièrement aux erreurs des systèmes et à leur analyse et éventuellement à l'utilisation de ce que nous connaissons de ces erreurs. Nous étudions dans ce travail les listes ordonnées d'erreurs générées à partir de différentes métriques et analysons ce qui en ressort. Nous avons appliqué la même méthode sur les sorties de différents systèmes de reconnaissance de la parole. Nos expériences mettent en évidence que certaines métriques apportent une information plus pertinente étant donné une tâche et transverse à différents systèmes.

ABSTRACT

Comparing error lists for ASR systems : contribution of different metrics.

The work presented here is concerned by the evaluation of automatic recognition systems used as a first step in a broader task on named entity recognition in spoken data. More precisely, the question we are asking is "what can an evaluation metric bring besides a score ?. We are in particular interested in the errors done by the systems and their analysis, and possibly the use of what we can learn of these errors. In this work we produced ordered error lists based on various metrics and analysed them. Our experiments show that some metrics provide a more pertinent information for the NER task that is almost system-independent.

MOTS-CLÉS : Reconnaissance automatique de la parole, Métriques d'évaluation, Analyse d'erreurs.

KEYWORDS: Automatic speech recognition, Metrics, Error analysis.

1 Introduction

Le travail que nous présentons ici [1] s'inscrit dans le domaine de l'évaluation des systèmes de reconnaissance automatique de la parole (RAP). Nous souhaitons évaluer ces systèmes sachant qu'ils seront utilisés dans une tâche aval et adapter l'évaluation de la transcription à ce type d'usage. Au vu des données à notre disposition, nous avons choisi ici comme tâche aval la reconnaissance des entités nommées (REN). Plus largement, nous nous interrogeons sur l'apport d'une métrique d'évaluation au

1. Cet article est une adaptation et extension de l'abstract soumis à LREC 2016.

delà de l'optimisation du score. Nous nous intéressons particulièrement aux erreurs des systèmes et à leur analyse, et éventuellement à l'utilisation de ce que nous connaissons de ces erreurs.

Les erreurs produites par les systèmes de RAP peuvent avoir un impact plus ou moins important selon leur fréquence, mais aussi selon l'utilisation qui est faite de la sortie du système (voir par exemple (Comas & Turmo, 2009) pour des systèmes de réponse précise à des questions en langue naturelle ou encore (Dinarelli & Rosset, 2011) pour des systèmes de REN). Ainsi comprendre et éventuellement anticiper les types des erreurs ou au moins leur impact sur la tâche qui suit celle de la RAP est un élément important pour améliorer la robustesse de ces systèmes, tant de RAP que de REN.

Plusieurs travaux ont porté sur l'analyse des erreurs produites par les systèmes de RAP. Ils avaient comme objectifs soit d'améliorer les systèmes de RAP eux-mêmes (Boháč *et al.*, 2012; Dufour & Esteve, 2008) soit de détecter automatiquement les erreurs (Ghannay *et al.*, 2015). Il est à noter que tenir compte des résultats de ces travaux restait à la charge des systèmes appliqués en aval.

Une grande partie des travaux qui se sont intéressés à l'étude des erreurs résiduelles des systèmes de RAP l'ont principalement fait dans le cadre de comparaisons entre erreurs produites par des systèmes et erreurs produites par des humains (Scharenborg, 2007; Lippmann, 1997). Ces études mettent en évidence que si les systèmes sont devenus plutôt très performants, ils ne sont pas encore en mesure de prendre en compte toutes les variations acoustiques observées. Par ailleurs, ces résultats montrent que les humains lorsqu'ils transcrivent (ou écoutent) de la parole obtiennent des performances cinq à six fois supérieures à celles des systèmes (Vasilescu *et al.*, 2012). Ces travaux ont permis de faire émerger des taxonomies d'erreurs indiquant que certains mots sont plus sujets à erreur que d'autres, en particulier les mots courts, pauvres d'un point de vue acoustique, ou encore les homophones courts potentiellement ambigus comme le verbe *a* et la préposition *à* (Adda-Decker, 2006). D'autres travaux se sont penchés sur la classification des erreurs. Par exemple, (Goryainova *et al.*, 2014) ont étudié les erreurs étant donné des classes morpho-syntaxiques de mots et (Santiago *et al.*, 2015) ont proposé une typologie d'erreurs fondée sur des critères syntaxiques et prosodiques ou encore (Rena Nemoto & Adda-Decker, 2008)) comparent des classifications fondées sur des études perceptives et des classifications automatiques des erreurs fréquentes sur des homophones.

Notre hypothèse de travail est qu'une métrique d'évaluation peut permettre de donner des indications sur les erreurs produites par un système de RAP. Plus précisément, nous considérons qu'une métrique doit pouvoir fournir une information sur la gravité d'une erreur, au sens de l'importance de celle-ci étant donnée la tâche globale. Cette gravité peut être dépendante de différents critères comme la fréquence certes, mais aussi son impact sur la capacité des systèmes intervenant en aval de la reconnaissance de la parole à fournir le meilleur résultat possible.

La section 2 présente notre proposition en y associant une brève description de différentes métriques utilisées pour l'évaluation des systèmes de RAP en contexte de REN. La section 3 présente les expériences que nous avons réalisées, les résultats obtenus et discute ceux-ci. Enfin la dernière section présente un bilan de cette expérience ainsi que quelques perspectives ouvertes.

2 Proposition

Considérant qu'obtenir une liste des erreurs et de leur importance passe par l'utilisation d'une métrique d'évaluation, nous commençons dans cette section par décrire quelques unes des métriques existantes. Puis nous détaillons notre proposition.

2.1 Évaluation des systèmes de reconnaissance automatique de la parole

La métrique la plus utilisée est le taux d'erreur mots (WER pour *Word Error Rate* (Pallett, 2003). Cette métrique compte les erreurs et normalise ce décompte par la taille (en nombre de mots) de la référence. Il existe différents types d'erreur, la substitution, la suppression et l'insertion, qui sont déterminés par un alignement de Levenshtein (Levenshtein, 1966) entre la référence (transcription manuelle) et l'hypothèse (transcription automatique). Le WER est donc une métrique fondée sur une énumération simple d'erreurs qui au final considère toutes les erreurs comme étant d'égale importance. Avec cette métrique plus une erreur est fréquente plus elle est grave.

Lorsque la RAP est la première étape d'une tâche plus globale, certaines études ont montré que le WER n'était pas toujours la métrique offrant la meilleure corrélation entre les performances du système de RAP et les performances obtenues sur la tâche globale comme dans le cas de la REN et de la recherche d'information (Garofolo *et al.*, 2000), de la traduction automatique (He *et al.*, 2011) ou encore de la compréhension de la parole (Wang *et al.*, 2003).

Plusieurs alternatives au WER existent. (Miller, 1955) a développé une mesure de la perte d'information occasionnée par les erreurs de RAP. Cette mesure, appelée le RIL (*Relative Information Loss*), est fondée sur le principe d'information mutuelle et permet d'obtenir une mesure de la dépendance statistique entre le vocabulaire de la référence et celui de l'hypothèse. Elle est représentée en termes d'entropie de Shannon. Par la suite, (Morris *et al.*, 2004) ont introduit le WIL (*Word Information Loss*) qui est une approximation du RIL. (Morris *et al.*, 2004) et (McCowan *et al.*, 2004) ont montré que ces deux métriques, RIL et WIL, présentent un intérêt lorsque le taux d'erreur est supérieur à 50%. Toujours dans le but de mesurer la perte d'information, (McCowan *et al.*, 2004) ont proposé d'adapter les métriques standards utilisées en extraction d'information, la précision (P), le rappel (R) et la f-mesure (F). L'idée générale consiste à calculer le rappel et la précision au niveau des mots en s'appuyant sur l'alignement entre la référence et l'hypothèse tel qu'il est produit par le calcul du WER. Comme nous nous intéressons à la reconnaissance d'entités nommées sur des données de parole, nous retenons la proposition de (Garofolo *et al.*, 1999) qui ont proposé de calculer un WER mais limité aux mots de la référence qui sont présents dans une entité nommée (NE-WER, *Named Entity Word Error Rate*). Un inconvénient de cette métrique est qu'elle ignore les mots insérés ou substitués en dehors des entités nommées mais qui peuvent conduire à une fausse alarme (détection erronée d'une entité nommée). Plus récemment et toujours pour évaluer la RAP dans le contexte de la REN, la métrique ATENE a été proposée par (Ben Jannet *et al.*, 2015). Cette métrique est fondée sur un modèle probabiliste qui estime le risque qu'une erreur de RAP induise une erreur de REN. ATENE a obtenu de meilleures corrélations que le WER, le NE-WER ou encore les mesures de pertes d'information (WIL, et P, R, F) entre les performances obtenus par les systèmes de RAP et celles des systèmes de REN.

Si disposer d'une métrique permettant d'estimer la qualité d'un système de RAP en tâche aval est intéressant, cela ne permet toutefois pas d'obtenir une liste des erreurs les plus coûteuses ou les plus fréquentes. Une telle liste est cependant utile pour améliorer les systèmes de RAP (Dufour & Esteve, 2008). Notre objectif est donc de pouvoir produire une liste des erreurs avec leur fréquence mais aussi, et surtout, une classification de ces erreurs en fonction de leur coût étant donnée une tâche. Les mesures générales comme le RIL et le WIL ne permettent pas de quantifier l'impact d'erreurs spécifiques, elles donnent un point de vue global sur la qualité d'un système de RAP. Nous avons décidé de nous appuyer sur ATENE qui a montré une corrélation plus importante que le WER ou le triplet P, R et F pour ce travail (Ben Jannet *et al.*, 2015).

2.2 Listes d'erreurs

Pour générer une liste d'erreurs qui quantifie l'impact des erreurs selon la tâche de REN, la première étape est de générer une liste d'erreurs. La méthodologie utilisée pour calculer un WER, en produisant un alignement entre les mots de la références et ceux de l'hypothèse, produit une telle liste. Il s'agit d'ailleurs de la seule métrique permettant cela. Cet alignement classifie les erreurs en insertions, suppressions et substitutions. Dénombrer les erreurs de cette liste permet d'obtenir une liste d'erreurs ordonnées selon leur importance telle que considérée par la métrique WER.

Dans ce travail, nous avons décidé de conserver cette liste d'erreurs mais de calculer pour chaque erreur son poids en fonction de son impact sur la métrique ATENE. Pour chaque occurrence d'une erreur, la transformation correspondant à l'erreur réalisée dans la référence est appliquée pour créer une nouvelle hypothèse. Le score de cette hypothèse est calculé avec la métrique ATENE. Cette métrique se comportant comme un taux d'erreur, les scores de chaque occurence sont cumulés pour obtenir le poids final qui lui sera associé. Le score peut en pratique être obtenu par un petit nombre de calculs autour du site de chaque instance d'erreur.

Le résultat de ces calculs permet alors de trier la liste et d'ordonner les erreurs selon leur importance. L'outil d'évaluation *sclite* qui implémente le WER fournit une liste d'erreurs ordonnée selon leur fréquence dans le fichier *dtl* (liste WER). Nous avons créé une liste équivalente en utilisant le résultat de l'impact de l'erreur sur ATENE comme élément d'importance (liste Atene). Afin d'affiner nos comparaisons, nous avons également produit deux autres listes : la première, inspirée par le NE-WER, ne considère que les occurrences d'erreurs apparaissant dans une entité nommée (liste In) et une seconde ne prenant en compte que les erreurs dans ou au contact d'une entité nommée (liste Near).

3 Expérimentations et résultats

3.1 Données

Pour nos expérimentations, nous avons utilisé les données de tests de la campagne ETAPE (Galibert *et al.*, 2014). Ces données, présentées dans le tableau 1, consistent en 15 émissions radiophoniques différentes, transcrites manuellement et par cinq systèmes de RAP (les participants à la campagne) ainsi que par un système rover.

	Test
Mots	115 803
Entités	5 933

	RAP-1	RAP-2	RAP-3	RAP-4	RAP-5	ROVER
WER	22,3	25,7	26,6	30,4	36,7	28,68

TABLE 1 – Description du corpus ETAPE : quantité de données et performance des systèmes

3.2 Méthodologie

Nous avons généré les listes des erreurs les plus impactantes selon chacune des métriques (ATENE, WER, In et Near) pour chaque système. Ces listes ont été générées avec deux seuils différents : la

Liste-10 conserve pour chaque système et chaque métrique les 10 erreurs les plus graves et inscrit pour chacune de ces erreurs le rang obtenu pour chaque métrique, la Liste-100 fait de même avec les 100 erreurs les plus graves. Si aucun recouvrement d'erreur n'existait entre les systèmes et les métriques, c'est à dire si chaque système et chaque métrique donnait des erreurs différentes, alors la fusion des Liste-10 devrait contenir $10\,(erreurs) \times 6\,(systèmes) \times 4\,(métriques) = 240$ entrées. Pourtant, avec les recouvrements, la fusion des Liste-10 se composent de 47 entrées. De la même manière, pour Liste-100, le nombre maximal d'entrées est de $100 \times 6 \times 4 = 2400$ entrées. Nous en observons en réalité 733. Ceci montre déjà que le recouvrement est important.

Nous souhaitons comparer ces listes afin de vérifier les hypothèses suivantes :

1. les listes générées par les quatre métriques pour un même système sont différentes. Si cette hypothèse est vérifiée, cela signifie que l'impact mesuré des erreurs d'un même système n'est pas le même selon la métrique utilisée ;
2. l'ordre d'importance des erreurs proposé par une métrique est équivalent quel que soit le système. Si cette hypothèse est vérifiée, cela signifie que la métrique est cohérente.

Pour vérifier ces deux hypothèses nous utilisons des mesures de corrélations de rang de Spearman (Spearman, 1904). Cette mesure donne des valeurs comprises entre -1 et +1 indiquant la puissance de corrélation entre les deux variable testées. Si la valeur est élevée ($\geq 0,8$), cela signifie que l'ordre des erreurs est le même. Au contraire, si la valeur absolue de la corrélation de rang est basse, cela signifie que les listes sont différentes et que ce ne sont pas les mêmes erreurs qui sont considérées comme importantes. Une valeur très négative indique que les listes sont en ordre inverse.

Une autre façon d'estimer la qualité de ces listes est de comparer leur contenu et les rangs des erreurs relevés, notamment en rapport avec la tâche. Si cette analyse ne peut être que partielle, elle peut donner des indications intéressantes.

3.3 Comparaisons de liste

Les corrélations de rang ont été calculées deux à deux pour chaque système et présentées sous forme de matrices sur les figures 1 et 2 respectivement pour Liste-10 et Liste-100. Les matrices se lisent de la manière suivante : plus une corrélation est forte ($R > 0,8$), plus la case est verte. Plus une corrélation est faible ($R < 0,3$), plus la case est rouge.

Analysons dans un premier temps la cohérence des métriques en fonction des systèmes afin de vérifier si pour une métrique donnée, les listes générées sont équivalentes pour tous les systèmes. La moyenne de corrélation de WER est respectivement de 0,96 (σ = 0,04) et de 0,91 (σ = 0,06) pour Liste-10 et Liste-100. Pour ATENE, la moyenne de corrélation est de 0,85 (σ = 0,10) et de 0,83 (σ = 0,09) Liste-10 et Liste-100. Pour Near, si pour Liste-10, la corrélation moyenne est de 0,84 (σ = 0,14), elle descend à 0,72 (σ = 0,16) sur Liste-100. Enfin, pour In, la moyenne des corrélations est de 0,66 (σ = 0,22) et de 0,64 (σ = 0,23) pour Liste-10 et Liste-100. Nous observons donc que les listes WER et Atene sont les plus cohérentes entre systèmes avec une corrélation moyenne supérieure à 0,8. En revanche la métrique In est celle qui semble la moins cohérentes avec une corrélation inférieure à 0,7. Les métriques WER et Atene quantifient donc les erreurs de manière très ressemblante d'un système à l'autre, et une prise en compte de la liste d'erreur d'un système RAP donné dans le système REN devrait être robuste au changement de système RAP.

Notre seconde question est de savoir si les listes d'erreurs sont différentes selon les métriques afin d'estimer leur complémentarité. Concernant le WER, pour Liste-10, sa corrélation moyenne avec

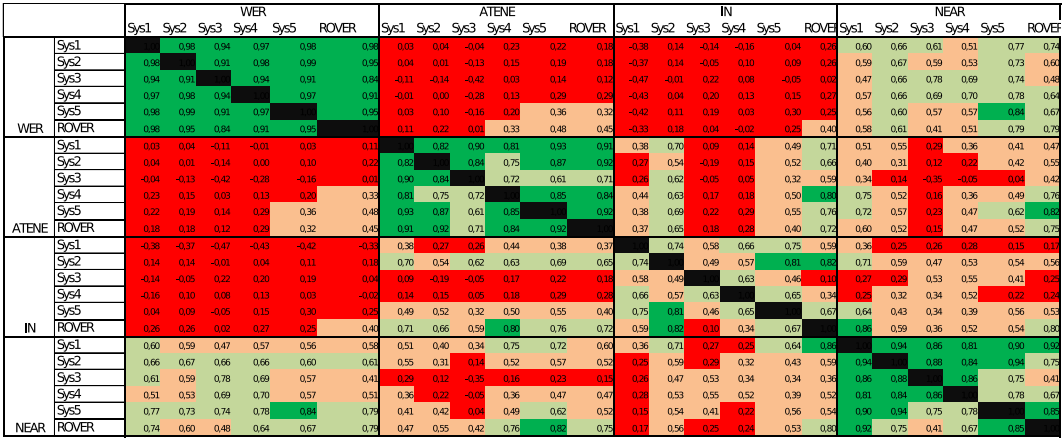


FIGURE 1 – Matrice de corrélation des mesures WER, ATENE, IN et Near pour les 5 systèmes de RAP et le ROVER sur Liste-10

ATENE est de 0,10 (σ = 0,19), de 0,03 (σ = 0,23) avec In et de 0,64 (σ = 0,10) avec Near. La même tendance s'observe avec Liste-100 : sa corrélation moyenne est de -0,17 (σ = 0,10) avec ATENE, de 0,14 (σ = 0,15) avec In et de 0,57 (σ = 0,09) avec Near. La corrélation très faible entre WER et In indique que les mots à l'intérieur des entités sont vraiment spécifiques par rapport à la langue globale. En revanche, WER corrèle mieux avec une métrique qui met en avant les erreurs autour des EN comme Near. Ceci est vraisemblablement dû au fait d'ajouter les mots autour des EN dilue leur spécificité en se rapprochant de la liste globale de WER.

Si on observe les corrélations entre ATENE et d'autres métriques, on constate une très faible corrélation entre WER et ATENE (0,10) montrant que ces deux mesures mettent en avant des erreurs différentes. Sa corrélation moyenne est un peu plus forte avec Near (0,40, σ = 0,11) et In (0,41, σ = 0,25) pour Liste-10. ATENE semble donc bien fournir des informations très différentes du WER grâce à sa prise en compte de la REN, le rapprochant de In mais allant plus loin.

3.4 Analyse qualitative

Voyons à présent plus en détail les erreurs jugées comme impactantes par ces différentes métriques.

Le tableau 2 contient quelques exemples extraits des différentes listes. Pour chacun des types d'erreur (suppressions, insertions et substitutions) il inclut l'erreur la plus importante pour chaque liste [2].

Comme nous pouvons le voir, la suppression considérée comme étant la plus importante par ATENE concerne la préposition *à*. Cette préposition est en général un bon indice pour la REN et nous pouvons voir que dans la référence ETAPE, 68% des occurrences de *à* se retrouvent devant une entité nommée. Il est probable que la suppression d'un tel mot a une conséquence sur un système de REN. La suppression de ce mot est également relativement importante pour le WER, ne serait-ce qu'à cause de sa fréquence (209 instances, 1,1% des erreurs). En revanche, nous pouvons constater que cette

2. Les listes d'erreurs ordonnent toutes les erreurs quel que soit leur type. Ainsi la première erreur de type X peut se retrouver au-delà de la première position dans la liste générale.

		WER						ATENE						IN						NEAR					
		Sys1	Sys2	Sys3	Sys4	Sys5	ROVER	Sys1	Sys2	Sys3	Sys4	Sys5	ROVER	Sys1	Sys2	Sys3	Sys4	Sys5	ROVER	Sys1	Sys2	Sys3	Sys4	Sys5	ROVER
	Sys1		0,96	0,94	0,95	0,96	0,88	-0,37	-0,08	-0,25	-0,08	-0,03	0,02	-0,02	0,09	0,18	0,01	0,23	0,27	0,63	0,63	0,62	0,55	0,66	0,52
	Sys2	0,96		0,93	0,93	0,92	0,86	-0,15	-0,38	-0,25	-0,19	-0,16	-0,06	-0,08	0,20	0,25	0,02	0,21	0,23	0,53	0,73	0,66	0,59	0,68	0,52
	Sys3	0,94	0,93		0,90	0,87	0,76	-0,30	-0,25	-0,34	-0,25	-0,17	-0,12	-0,13	-0,02	0,25	0,08	0,12	0,18	0,54	0,61	0,73	0,62	0,58	0,43
	Sys4	0,95	0,93	0,90		0,87	0,77	-0,17	-0,25	-0,30	-0,37	-0,16	-0,05	-0,13	-0,10	0,13	0,23	0,08	0,14	0,50	0,54	0,60	0,71	0,54	0,43
	Sys5	0,96	0,92	0,87	0,87		0,92	-0,09	-0,22	-0,19	-0,15	-0,19	-0,08	-0,06	0,01	0,19	0,10	0,40	0,29	0,49	0,54	0,56	0,53	0,71	0,53
WER	ROVER	0,88	0,86	0,76	0,77	0,92		-0,03	-0,12	-0,11	-0,03	-0,03	-0,03	0,11	0,10	0,28	0,19	0,40	0,47	0,46	0,48	0,45	0,44	0,67	0,67
	Sys1	-0,37	-0,15	-0,30	-0,17	-0,09	-0,03		0,88	0,82	0,82	0,86	0,84	0,30	0,31	0,12	0,26	0,35	0,39	0,02	0,08	-0,10	0,05	0,17	0,30
	Sys2	-0,08	-0,38	-0,25	-0,25	-0,22	-0,12	0,88		0,74	0,77	0,78	0,80	0,32	0,14	0,06	0,13	0,22	0,29	0,22	-0,10	-0,09	-0,01	0,05	0,18
	Sys3	-0,25	-0,25	-0,34	-0,30	-0,19	-0,11	0,82	0,74		0,78	0,73	0,73	0,39	0,32	0,10	0,15	0,30	0,40	0,10	0,04	-0,18	-0,10	0,04	0,13
	Sys4	-0,08	-0,19	-0,25	-0,37	-0,15	-0,03	0,82	0,77	0,78		0,73	0,74	0,46	0,34	0,18	0,18	0,35	0,33	0,21	0,05	-0,04	-0,08	0,12	0,17
	Sys5	-0,03	-0,16	-0,17	-0,16	-0,19	-0,03	0,86	0,78	0,73	0,73		0,88	0,42	0,28	0,22	0,24	0,36	0,37	0,21	0,06	-0,02	0,00	0,12	0,21
ATENE	ROVER	0,02	-0,06	-0,12	-0,05	-0,08	-0,03	0,84	0,80	0,73	0,74	0,88		0,44	0,31	0,18	0,28	0,38	0,41	0,24	0,11	0,01	0,06	0,21	0,25
	Sys1	-0,02	-0,08	-0,13	-0,13	-0,06	0,11	0,30	0,32	0,39	0,46	0,42	0,44		0,61	0,56	0,69	0,70	0,73	0,39	0,17	0,08	0,16	0,29	0,40
	Sys2	0,09	0,20	-0,02	-0,10	0,01	0,10	0,31	0,14	0,32	0,34	0,28	0,31	0,61		0,53	0,49	0,54	0,51	0,41	0,44	0,16	0,12	0,27	0,39
	Sys3	0,18	0,25	0,25	0,13	0,19	0,28	0,12	0,06	0,10	0,18	0,22	0,18	0,56	0,53		0,47	0,51	0,57	0,47	0,39	0,46	0,31	0,35	0,38
	Sys4	0,01	0,02	0,08	0,23	0,10	0,19	0,26	0,13	0,15	0,18	0,24	0,28	0,69	0,49	0,47		0,39	0,44	0,38	0,22	0,25	0,53	0,29	0,37
	Sys5	0,23	0,21	0,12	0,08	0,40	0,40	0,35	0,22	0,30	0,35	0,36	0,38	0,70	0,54	0,51	0,39		0,77	0,46	0,31	0,27	0,18	0,63	0,58
IN	ROVER	0,27	0,23	0,18	0,14	0,29	0,47	0,39	0,29	0,40	0,33	0,37	0,41	0,73	0,51	0,57	0,44	0,77		0,49	0,36	0,32	0,29	0,50	0,70
	Sys1	0,63	0,53	0,54	0,50	0,49	0,46	0,02	0,22	0,10	0,21	0,21	0,24	0,39	0,41	0,47	0,38	0,46	0,49		0,78	0,76	0,73	0,75	0,60
	Sys2	0,63	0,73	0,61	0,54	0,54	0,48	0,08	-0,10	0,04	0,05	0,06	0,11	0,17	0,44	0,39	0,22	0,31	0,36	0,78		0,77	0,70	0,72	0,56
	Sys3	0,62	0,66	0,73	0,60	0,56	0,45	-0,10	-0,09	-0,18	-0,04	-0,02	0,01	0,08	0,16	0,46	0,25	0,27	0,32	0,76	0,77		0,75	0,66	0,46
	Sys4	0,55	0,59	0,62	0,71	0,53	0,44	0,05	-0,01	-0,10	-0,08	0,00	0,06	0,16	0,12	0,31	0,53	0,18	0,29	0,73	0,70	0,75		0,57	0,46
	Sys5	0,66	0,68	0,58	0,54	0,71	0,67	0,17	0,05	0,04	0,12	0,12	0,21	0,29	0,27	0,35	0,29	0,63	0,50	0,75	0,72	0,66	0,57		0,79
NEAR	ROVER	0,52	0,52	0,43	0,43	0,53	0,67	0,30	0,18	0,13	0,17	0,21	0,25	0,40	0,39	0,38	0,37	0,58	0,70	0,60	0,56	0,46	0,46	0,79	

FIGURE 2 – Matrice de corrélation des mesures WER, ATENE, In et Near pour les 5 systèmes de RAP et le ROVER sur Liste-100

Suppressions				
Atene	Mot	WER	In	Near
1	à	7	16	57
873	il	**1**	9	7
9	de	5	**1**	**1**

Insertions				
Atene	Mot	WER	In	Near
10	dix	673	-	2220
8112	et	**14**	16	**4**
8184	des	87	**12**	23

Substitutions					
Atene	Ref	Hyp	WER	In	Near
2	de	deux	124	98	48
4863	il	qui	**38**	-	-
8	deux	de	116	**4**	**18**

TABLE 2 – Extrait des listes générées, avec le rang associé à chacune des métriques, Atene, WER, les erreurs dans les entités (In) et les erreurs autour des entités (Near). En gras apparaît le rang dans chaque liste de la première erreur de chaque type.

suppression se trouve sur un rang relativement bas pour les deux autres listes. L'importance de cette erreur ne ressort pas particulièrement pour les listes In et Near (respectivement rang 16 et 57).

L'insertion la plus importante relevée par Atene est celle du nombre *dix*. Ce mot se retrouve, dans le corpus de référence, toujours dans une entité (Montant, Date ou Heure). Les trois autres métriques ne la considèrent pas importante (rang 673 pour le WER, inexistant pour In et 2220 pour Near).

La première substitution, le *de* (préposition ou partitif) substitué par *deux* (nombre) présente un peu le même problème avec les listes fondées sur les métriques traditionnelles. Cette substitution présente un risque élevé de soit provoquer une fausse alarme (détection erronée d'une entité) soit de segmenter une entité longue. Mais elle n'est pas considérée comme importante par les autres métriques (rang 124 pour le WER, 98 pour In et 48 pour Near). À l'inverse la substitution inverse de *deux* vers *de* est considérée comme plus importante par les méthodes In et Near (rang 4 et 18).

La suppression la plus importante pour le WER est celle du pronom *il* et les insertions les plus

importantes sont celles de la conjonction de coordination *et* et du déterminant *des*. Ces erreurs ne semblent pas pouvoir avoir un impact fort sur la REN, où qu'elles se produisent. Mais elles sont fréquentes, probablement car ces mots sont monophoniques, et sont du coup considérées haut placées.

La supression de *de* est considérée comme importante quelle que soit la méthode utilisée pour générer la liste d'erreurs. Ceci n'a rien d'étonnant puisque il s'agit d'une suppression fréquente et qu'en plus ce mot peut servir de marqueur pour la REN.

4 Conclusions et perspectives

Nous avons présenté une méthode pour générer des listes d'erreurs produites par une système de RAP ordonnées selon leur impact sur une tâche de détection d'entités nommées. Cette méthode s'appuie d'une part sur l'alignement tel que généré par le calcul de la métrique WER pour générer la liste initiale et la méthodologie liée à la métrique ATENE pour l'estimation de l'impact des erreurs.

Nous avons appliqué cette méthode aux données de test de la campagne ETAPE et avons comparé la liste ordonnée telle que générée par notre méthode avec des listes générées par d'autres méthodes. Les mesures de corrélation montrent que les métriques WER et ATENE sont cohérentes pour chaque système et fournissent une information en grande partie indépendante du système. Elles sont de plus assez peu corrélées entre elles, donnant des informations différentes sur l'importance des erreurs.

Une analyse plus détaillée des listes a ensuite permis de montrer que la quantification d'impact via ATENE fournit bien une information importante pour la tâche REN et pas uniquement fréquencielle. Nous avons donc bien réussi a établir une liste ordonnée d'erreurs motivée par le système aval.

L'étape suivante sera bien évidemment d'exploiter cette liste. Corriger le système de RAP n'est probablement pas possible, les erreurs importantes sont en l'occurence des erreurs difficiles. Nous essairons donc de prendre le problème dans l'autre sens et de rendre le système REN plus robuste face aux erreurs identifiées.

Remerciements

Ce travail a été financé partiellement par le projet VERA - ANR 12 BS02 006 04.

Références

ADDA-DECKER M. (2006). De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. In *Proc of JEP*, Dinard, France.

BEN JANNET M. A., GALIBERT O., ADDA-DECKER M. & ROSSET S. (2015). How to evaluate asr output for named entity recognition ? In *Interspeech*, Dresden, Germany.

BOHÁČ M., NOUZA J. & BLAVKA K. (2012). Investigation on most frequent errors in large-scale speech recognition applications. In *Text, Speech and Dialogue*, p. 520–527 : Springer.

COMAS P. R. & TURMO J. (2009). Robust question answering for speech transcripts : Upc experience in qast 2009. In *Working Notes of CLEF 2009*, Corfou, Grèce.

DINARELLI M. & ROSSET S. (2011). Models Cascade for Tree-Structured Named Entity Detection. In *IJCNLP*, p. 1269–1278, Chiang Mai, Thailand.

DUFOUR R. & ESTEVE Y. (2008). Correcting asr outputs : Specific solutions to specific errors in french. In *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*, p. 213–216.

GALIBERT O., LEIXA J., ADDA G., CHOUKRI K. & GRAVIER G. (2014). The ETAPE speech processing evaluation. In *LREC*, Reykjavik, Iceland.

GAROFOLO J. S., AUZANNE C. G. & VOORHEES E. M. (2000). The trec spoken document retrieval track : A success story. *NIST SPECIAL PUBLICATION SP*, **500**(246), 107–130.

GAROFOLO J. S., VOORHEES E. M., AUZANNE C. G., STANFORD V. M. & LUND B. A. (1999). 1998 trec-7 spoken document retrieval track overview and results. In *Broadcast News Workshop'99 Proceedings*, p. 215 : Morgan Kaufmann Pub.

GHANNAY S., ESTÈVE Y. & CAMELIN N. (2015). Word embeddings combination and neural networks for robustness in asr error detection. In *EUSIPCO*, Nice, France.

GORYAINOVA M., GROUIN C., ROSSET S. & VASILESCU I. (2014). Morpho-syntactic study of errors from speech recognition system. In *LREC*, Reykjavik, Iceland.

HE X., DENG L. & ACERO A. (2011). Why word error rate is not a good metric for speech recognizer training for the speech translation task ? In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, p. 5632–5635 : IEEE.

LEVENSHTEIN V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, **10**(8), 707–710.

LIPPMANN R. P. (1997). Speech recognition by machines and humans. *Speech Communication*, **22**(1), 1–15.

MCCOWAN I. A., MOORE D., DINES J., GATICA-PEREZ D., FLYNN M., WELLNER P. & BOURLARD H. (2004). *On the use of information retrieval measures for speech recognition evaluation*. Rapport interne, IDIAP.

MILLER G. A. (1955). Note on the bias of information estimates. *Information theory in psychology : Problems and methods*, **2**, 95–100.

MORRIS A. C., MAIER V. & GREEN P. (2004). From wer and ril to mer and wil : improved evaluation measures for connected speech recognition. In *INTERSPEECH*.

PALLETT D. S. (2003). A look at nist's benchmark asr tests : past, present, and future. In *ASRU'03*.

RENA NEMOTO I. V. & ADDA-DECKER M. (2008). Speech errors on frequently observed homophones in french : Perceptual evaluation vs automatic classification. In *LREC*, Marrakech, Morocco.

SANTIAGO F., ADDA-DECKER M. & DUTREY C. (2015). Towards a typology of asr errors via syntax-prosody mapping. In *Errare Workshop*, Sinaia, Romania.

SCHARENBORG O. (2007). Reaching over the gap : A review of efforts to link human and automatic speech recognition research. *Speech Communication*, **49**(5), 336–347.

SPEARMAN C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, **15**, 72–101.

VASILESCU I., ADDA-DECKER M. & LAMEL L. (2012). Cross-lingual studies of ASR errors : paradigms for perceptual evaluations. In *LREC*, Istanbul, Turkey.

WANG Y.-Y., ACERO A. & CHELBA C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. In *ASRU'03*, p. 577–582 : IEEE.

Se concentrer sur les différences : une méthode d'évaluation subjective efficace pour la comparaison de systèmes de synthèse

Jonathan Chevelu[1] Damien Lolive[1] Sébastien Le Maguer[2] David Guennec[1]
(1) IRISA, Université de Rennes 1, Lannion, France
(2) Saarland University, Saarbrücken, Germany
jonathan.chevelu@irisa.fr, damien.lolive@irisa.fr,
david.guennec@irisa.fr, slemaguer@coli.uni-saarland.de

RÉSUMÉ

En proposant une nouvelle approche de synthèse de la parole, les études comportent généralement une évaluation subjective d'échantillons acoustiques produits par un système de référence et un nouveau système. Ces échantillons sont produits à partir d'un petit ensemble de phrases choisies aléatoirement dans un unique domaine. Ainsi, statistiquement, des échantillons pratiquement identiques sont présentés et réduisent les écarts de mesure entre les systèmes, au risque de les considérer comme non significatifs. Pour éviter cette problématique méthodologique, nous comparons deux systèmes sur des milliers d'échantillons de différents domaines. L'évaluation est réalisée uniquement sur les paires d'échantillons les plus pertinentes, c'est-à-dire les plus différentes acoustiquement. Cette méthode est appliquée sur un système de synthèse de type *HTS* et un second par sélection d'unités. La comparaison avec l'approche classique montre que cette méthode révèle des écarts qui jusqu'alors n'étaient pas significatifs.

ABSTRACT

Focus on differences : a subjective evaluation method to efficiently compare TTS systems *.

When trying to assess the effectiveness of a new speech synthesis method, researchers usually conduct subjective evaluations by randomly choosing a small set of samples, from the same domain, taken from a baseline system and the proposed one. When selecting them randomly, statistically, samples with almost no differences are evaluated and the global measure is smoothed which may lead to judge the improvement not significant. To solve this methodological flaw, we propose to compare speech synthesis systems on thousands of generated samples from various domains and to focus subjective evaluations on the most relevant ones by computing a normalized alignment cost between sample pairs. This process has been successfully applied both in the HTS statistical framework and in the unit selection approach. A comparison between tests involving most different samples and randomly chosen samples shows clearly that the proposed approach reveals significant differences between the systems.

MOTS-CLÉS : synthèse de la parole, évaluation subjective, différence acoustique.

KEYWORDS: speech synthesis, subjective evaluation, acoustic difference.

∗. Cet article reprend un travail présenté par les mêmes auteurs à la conférence Interspeech 2015.

1 Introduction

Dans le domaine de la synthèse de la parole (TTS), l'évaluation subjective est cruciale puisque l'objectif principal est de produire un message destiné à des auditeurs humains. Des évaluations objectives et subjectives peuvent être utilisées. D'une part, les évaluations objectives ont l'avantage d'être peu coûteuses à réaliser, mais peu importe leur précision, elles ne peuvent pas encore remplacer les tests subjectifs. D'autre part, pour être intéressantes, les évaluations subjectives ont besoin d'un grand nombre d'auditeurs et d'un grand nombre d'échantillons choisis en fonction du domaine du contexte d'utilisation de la synthèse.

Plusieurs types d'évaluations perceptives sont généralement utilisées. Parmi toutes ces méthodes, on peut distinguer des tests de préférence comme AB et ABX, des tests de pointage comme MOS (*Mean Opinion Score*), DMOS (*Degradation MOS*) ou plus récemment MUSHRA (*MUltiple Stimuli with Hidden Reference and Anchor*). Toutes ces méthodes ont le même objectif, à savoir le classement des systèmes selon certains critères subjectifs.

Dans la littérature, la plupart des propositions scientifiques sont évaluées à l'aide de tests perceptifs mais le nombre d'échantillons étudiés reste très limité. Par exemple, le défi Blizzard est composé de campagnes d'évaluation à grande échelle (King & Karaiskos, 2012; Prahallad *et al.*, 2014) mais ne comporte que quelques centaines de signaux. Ceci se retrouve dans d'autre travaux, parmi lesquels nous pouvons citer (Sainz *et al.*, 2014) avec 350 phrases, (Garcia *et al.*, 2006) avec 7 phrases pour 5 systèmes ou encore (Hinterleitner *et al.*, 2011) avec deux groupes de 18 stimuli. Ce faible nombre de stimuli est généralement motivé par l'aspect particulièrement chronophage des campagnes d'évaluation perceptives. Quelques travaux récents ont mis en doute la méthodologie d'évaluation, comme (Latorre *et al.*, 2014) qui étudie l'impact des références mentales des auditeurs sur les résultats des tests perceptifs, ou (Hinterleitner *et al.*, 2011; Viswanathan & Viswanathan, 2005) qui ont proposé des modifications des protocoles existants. Des alternatives aux méthodes classiques ont également été utilisées, sur le principe d'évaluations en ligne à grande échelle (*crowdsourcing*) comme décrit dans (Buchholz *et al.*, 2013).

Plus important que le petit nombre d'échantillons choisis, le fait qu'ils soient choisis au hasard et non pas pour leur importance pour les méthodes d'évaluation peut biaiser les résultats des évaluations. Dans cet article, contrairement à ce qui se fait habituellement, nous proposons de synthétiser un grand nombre d'échantillons (plusieurs milliers), à partir de textes de divers domaines. Compte tenu du nombre élevé d'échantillons, nous introduisons un coût d'alignement entre les paires d'échantillons provenant de deux systèmes différents afin de les classer par similarité acoustique. Une fois cela fait, nous pouvons construire une évaluation perceptive en utilisant uniquement les signaux les plus différents. De cette façon, nous ne faisons aucune hypothèse concernant la qualité d'un système parmi les autres, nous concentrons simplement l'évaluation sur ce qui peut faire la différence entre les systèmes. Une telle stratégie permet de réduire la taille d'une évaluation perceptive pour recentrer l'évaluation sur les différences importantes entre les systèmes. Cette méthode a été utilisée avec succès à la fois avec une paire de systèmes statistique (*HTS*) et une paire de système par sélection d'unités. Les résultats que nous obtenons pour des tests de préférence AB sont clairement significatifs, alors que lorsque les phrases à vocaliser sont sélectionnées aléatoirement, les écarts de performance observés n'arrivaient pas à discriminer les systèmes.

Le reste de l'article est organisé comme suit. Dans la section 3, nous présentons les systèmes que nous utilisons dans les expériences. La section 4 décrit la méthodologie pour construire les évaluations. Enfin, la section 5 présente les expériences ainsi que les résultats.

2 Corpus de parole

Pour les besoins de cette étude, deux corpus de parole sont utilisés. Le premier corpus est extrait à l'aide d'un processus entièrement automatique présentée dans (Boeffard *et al.*, 2012), à partir d'un livre audio en français. Le locuteur masculin réalise une lecture moyennement expressive et le signal est échantillonné à $44,1$ kHz. Le corpus annoté complet contient 3339 énoncés (10h45 de parole). Pour les expériences, 1h de parole a été extraite du corps pour former le système de synthèse à base de HMM, décrit plus loin. Par la suite, ce corpus sera appelé *Audiobook*.

Le deuxième corpus est produit par un locuteur féminin en français. Il a été initialement construit pour le système de synthèse d'un serveur vocal interactif utilisé par un opérateur de télécommunications. Ses annotations ont été contrôlées manuellement. Le corpus complet contient 7h de parole enregistrées à 16 kHz. Dans la suite, ce corpus est appelé *SVI*.

3 Systèmes de synthèses de la parole

Afin d'évaluer l'efficacité de la méthode proposée, deux systèmes de synthèse de la parole sont utilisés. Le premier est basée sur *HTS* et le second est un système de Synthèse Par Corpus (*SPC*).

3.1 Synthèse par HMM

Au cours de la dernière décennie, le système *HTS* a été largement popularisé et utilisé pour de nombreuses études. Ce système de synthèse fondé sur les modèles de Markov cachés (*HMM-Based*) s'est révélé être une méthode très flexible pour produire de la parole (Tokuda *et al.*, 2002; Zen *et al.*, 2009). Cette méthode statistique repose sur la structure de modèles semi-markoviens cachés pour modéliser les coefficients Mel-généralisés (MGC), l'apériodicité, la fréquence fondamentale ($F0$) comme des flux séparés et, à l'aide d'arbres de décision, les associer à un ensemble de descripteurs (Zen *et al.*, 2009). Nous avons utilisé la version 2.3 alpha d'*HTS* avec 50 coefficients MGC, 25 coefficients de bandes apériodicité (BAP) et la $F0$.

Dans cet article, nous nous concentrons volontairement sur deux ensembles de fonctionnalités simples constituées par les phonèmes étiquetés, y compris l'étiquette du phonème en cours et les étiquettes de contexte en utilisant soit des fenêtres [-1,1], soit des fenêtres [-2,2] (un ou deux phonèmes avant et un ou deux après le phonème sélectionné). Ces configurations sont choisies en suite aux travaux de (Le Maguer *et al.*, 2013) qui ont évalué l'importance des descripteurs du jeu standard proposé par *HTS* pour le français. Il se trouve que la taille de la fenêtre phonétique utilisée a été jugée comme l'un des critères les plus pertinents, mais sans arriver à observer un écart important lors de l'évaluation perceptive. Néanmoins, en appliquant la méthodologie que nous proposons, nous allons montrer qu'il existe bien une différence significative.

À partir du corpus *Audiobook*, nous avons entraîné deux systèmes *HTS* :

— *HMM-p3* : utilise uniquement comme descripteur les étiquettes du phonème courant considéré et celles du phonème précédent et suivant ;

— *HMM-p5* : utilise les descripteurs de *HMM-p3* auxquels s'ajoute les étiquettes des deux phonèmes encadrant ceux précédemment considérés.

3.2 Synthèse par sélection d'unités

3.2.1 Système de référence

Le système de synthèse par sélection d'unités utilisé dans cette étude est celui décrit dans (Guennec & Lolive, 2014). Le coût de concaténation que nous utilisons ici comporte les trois composantes que sont les distances sur les MFCC, l'amplitude et la $F0$ entre deux unités. Pour accélérer le processus de sélection, une étape de présélection similaire à celle proposé dans (Conkie *et al.*, 2000) est employée pour filtrer les unités candidates. Les filtres utilisés agissent comme un coût cible binaire et la fonction de coût à optimiser est réduite à un coût de concaténation. Nous supposons donc que deux unités candidates passant l'étape de présélection sont équivalentes en ce qui concerne le coût cible. Les filtres suivants sont utilisées dans le système de base :

— le phonème est-il dans la dernière syllabe d'une phrase ?
— le phonème est-il dans la dernière syllabe d'un groupe syntaxique ?
— le phonème est-il dans la dernière syllabe d'un mot ?
— le phonème est-il dans une syllabe à l'intonation montante ?

3.2.2 Système de comparaison

Dans le domaine de la synthèse de la parole, la réduction de corpus est un problème général largement étudié. Comme le montre la littérature, plusieurs articles ont étudié les moyens de réduire un corpus de parole ou de texte, afin de minimiser la durée d'enregistrement ou la taille des voix manipulées. En particulier, (Lambert *et al.*, 2007) propose une évaluation de l'impact de la réduction sur la qualité d'un système de synthèse. Il y est montré qu'un corpus sélectionné aléatoirement semble produire une qualité perçue semblable à celle qui a été obtenue par l'utilisation d'un corpus construit par simple couverture des diphones. Ce point particulier est étudié en utilisant la méthodologie que nous proposons.

Le problème de la réduction de corpus peut être considérée comme un problème de couverture d'ensemble (SCP) (François & Boeffard, 2002). Celui-ci étant un problème NP-difficile, la stratégie la plus fréquente utilisée pour le résoudre repose sur des algorithmes gloutons. Compte tenu de la répartition des attributs souhaités dans les corpus linguistiques, de nombreux types d'algorithmes gloutons ont été étudiés, par exemple dans (François & Boeffard, 2002) et (Krul *et al.*, 2007). À l'aide de la relaxation lagrangienne, (Chevelu *et al.*, 2008) montre qu'un algorithme glouton par agglomération suivi d'un algorithme glouton de type *cracheur* est proche de la solution optimale.

Pour évaluer notre méthodologie, nous proposons de réduire un corpus de parole (le corpus *Complet*) en utilisant deux méthodes :

— *TTSCouv* couvre au minimum une fois chaque paire successive de phonèmes pour chaque configuration de descripteurs existante. Les descripteurs retenus sont ceux utilisé dans le moteur de synthèse décrit en 3.2.1.
— *CompAléa* est construit en complétant aléatoirement une couverture des diphonèmes réalisée sur *Complet* ($\sim$ 300 phrases) jusqu'à atteindre le même nombre de phones que dans *TTSCouv*.

En utilisant le corpus *SVI*, dont les principales statistiques sont présentées dans le tableau 1, deux systèmes de synthèse par sélection d'unités sont alors construits : ils réalisent les synthèses en utilisant respectivement *TTSCouv* et *CompAléa*, et ils sont appelés par le nom de leur corpus associé, sans risque de confusion.

Sub-corpus	*Complet*	*TTSCouv*	*CompAléa*
Durée	7h06'12	3h11'15	3h04'19
Taille en phrases	7,662	3,238	3,350
Taille en étiquettes	259,684	112,324	112,324
Nombre d'étiquettes	34 phonèmes et 2 *Non Speech sound* (NSS)		
Nombre de diphonèmes	1,242		

TABLE 1: Statistiques principales des corpus utilisés.

4 Méthodologie d'évaluation

Dans cette section, la méthode d'évaluation proposée et le corpus de test utilisé sont présentés.

4.1 Approche

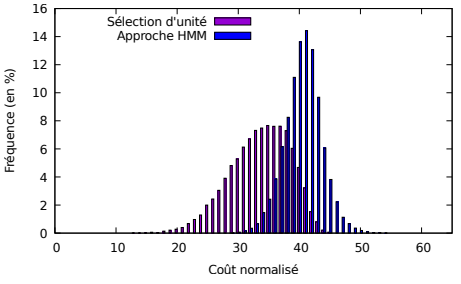
En général, l'approche classique pour les évaluations subjectives est de synthétiser un petit ensemble d'échantillons, de les proposer aux auditeurs, et de tirer des conclusions sur les systèmes à partir des résultats obtenus sur cette petite série d'échantillons. À notre avis, cette méthode fonctionne uniquement pour les systèmes qui ont une grande différence de qualité de sortie et dépend en grande partie de l'ensemble des phrases choisies. Pour révéler les différences entre deux systèmes, il faut cependant se concentrer sur les différences constatées dans les signaux de parole générés. Lorsque les évaluations se fondent sur un petit ensemble d'échantillons, les signaux les plus différents sont bien souvent absents. Par conséquent, nous proposons ce qui suit :

1. Synthétiser un grand nombre de textes provenant de domaines variés et de styles différents avec chaque système ;
2. Calculer pour chaque paire d'échantillons un coût d'alignement (par exemple une DTW – *Dynamic Time Warping* (Sakoe & Chiba, 1978)) ;
3. Sélectionner les échantillons les plus dissemblables pour évaluer les systèmes.

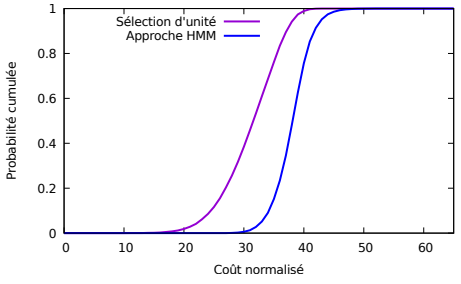
Dans le présent document, le coût d'alignement est calculé en fonction du coût de la DTW entre les vecteur de MFCC pour chaque signal, divisé par la longueur du chemin d'alignement. On obtient ainsi un coût normalisé. Cette mesure a l'avantage d'être indépendante des systèmes en cours d'évaluation mais d'autres peuvent être envisagées.

4.2 Corpus d'évaluation

Pour être indépendant des corpus de parole choisis, nous avons utilisé un corpus textuel provenant d'une source différente. Il est composé d'un ensemble de phrases extraites d'une collection de 50 livres électroniques couvrant de nombreux sujets et styles d'écriture. Les phrases qui en résultent sont ensuite filtrées pour garder celles comportant entre 30 et 60 phonèmes afin de produire des signaux d'une durée comprise approximativement entre 3 et 6 secondes (comme recommandé dans (ITU-T, 1996)). Puisque le même phonétiseur est utilisé pour les deux systèmes, les phrases avec des erreurs de phonétisation sont filtrées (généralement des phrases contenant des symboles non-standard ou des



(a) Histogrammes des valeurs des coûts.



(b) Fonctions de densité cumulée pour la mesure des coûts.

FIGURE 1: Distributions des coûts de DTW normalisés entre deux paires de systèmes évalués. Cette figure montre que la distribution a un comportement de type gaussien et qu'un nombre important d'échantillons semblent acoustiquement similaires.

noms propres). Enfin, parmi les phrases restantes, 27 030 sont extraites au hasard pour construire le corpus de test à synthétiser lors des expériences.

5 Expériences et résultats

5.1 Distribution des coûts d'alignement

La figure 1 montre la répartition des coûts de DTW normalisés pour les 27 030 phrases lorsque l'on compare *TTSCouv* et *CompAléa* (en rouge) et lorsque l'on compare *HMM-p3* et *HMM-p5* (en bleu). Considérant à la fois l'histogramme et la fonction de densité, on observe un comportement de type gaussien. La conséquence est que, lorsque des échantillons sont choisis aléatoirement, l'ensemble utilisé pour les évaluations perceptives peut contenir un nombre élevé d'échantillons acoustiquement équivalents. Ceci risque donc de lisser les résultats de l'évaluation perceptive et les systèmes peuvent être considérés comme équivalents.

Notons que les coûts pour les systèmes à base de HMM sont plus important en moyenne que ceux des systèmes par sélection d'unités. Ce résultat est prévisible puisque les deux systèmes par sélection sont construits à partir de la même voix, les signaux de sortie peuvent donc partager certains segments, ce qui n'est pas le cas pour les systèmes à base de HMM. Ainsi, malheureusement, il semble difficile de trouver un seuil universel sur le coût de DTW à partir duquel on pourrait dire que deux signaux sont significativement différents.

5.2 Évaluations perceptives

Afin d'évaluer la méthodologie proposée, nous avons mené des évaluations séparées pour les systèmes par sélection d'unités et ceux utilisant des HMM. Dans le premier cas, nous avons évalué trois méthodes d'échantillonnage. Le premier test consiste à sélectionner les échantillons de parole les plus similaires selon la mesure proposée et est fait pour vérifier que la mesure est corrélée à la perception

en termes de similarité. Le second est la méthode classique, à savoir sélectionnant de manière aléatoire un sous-ensemble d'échantillons. Enfin, le troisième test est basé sur la sélection des échantillons de parole les plus dissemblables. Dans le second cas, pour les systèmes à base de HMM, nous n'avons évalué que le sous-ensemble aléatoire d'échantillons et un sous-ensemble composé des échantillons vocaux les plus différents. Les statistiques de chaque corpus de test sont présentées dans le tableau 2. Elles montrent une différence significative entre les corpus de coûts maximaux et les autres.

Corpus de test	Nb. de phrases	Coût moyen (écart-type.)
Corpus de coûts min.	100	$15,0\ (1,6)$
Corpus aléatoire	100	$31,2\ (4,7)$
Corpus de coûts max.	100	$41,6\ (0,5)$
Corpus complet	27030	$31,2\ (4,9)$

(a) Corpus d'évaluation pour les systèmes par sélection d'unités.

Corpus de test	Nb. de phrases	Coût moyen (écart-type.)
Corpus aléatoire	100	$38,6\ (3,3)$
Corpus de coûts max.	100	$48,5\ (1,2)$
Corpus complet	27030	$34,0\ (3,0)$

(b) Corpus d'évaluation pour les systèmes par HMM.

TABLE 2: Statistiques des corpus d'évaluation.

Considérant les configurations énoncées précédemment, nous avons extrait 100 échantillons par système pour construire des tests de préférence de type AB. À chaque étape, deux signaux générés à partir de la même phrase, mais par des systèmes différents, sont présentés dans un ordre aléatoire. Il a été demandé à 10 auditeurs de choisir leur signal préféré (trois réponses sont proposées : A, B et *Indifférent*). Les résultats sont présentés dans les tableaux 3a et 3b.

Premièrement, nous pouvons observer que lors de la sélection des échantillons choisis aléatoirement, les systèmes ne sont pas différenciés et la préférence est uniformément répartie entre les trois choix possibles. Cela est vrai tant pour les systèmes par HMM que pour la sélection d'unités. En outre, la différence entre les systèmes n'est pas significative, selon un test binomial avec un intervalle de confiance à 95%. Cela peut s'expliquer par le fait que le choix aléatoire a tendance à sélectionner des échantillons contenant les événements les plus fréquents. On peut en outre supposer que, sur les événements les plus fréquents, deux systèmes proches peuvent se comporter de la même manière.

En sélectionnant les échantillons via la méthode de classement proposée dans le présent article et en ne gardant que les plus dissemblables pour l'évaluation, les résultats montrent clairement une préférence pour un système. À chaque fois, les systèmes que l'on pouvait considérer comme probablement meilleurs (*TTSCouv* et *HMM-p5*) obtiennent effectivement les meilleurs résultats. En outre, dans les deux cas, le nombre de réponses de type *Indifférent* diminue considérablement (par exemple, il est divisé par 2 pour les systèmes par sélection d'unités). Pour les deux systèmes, les résultats des tests perceptifs sont maintenant significatifs. Par conséquent, le classement proposé à permis de concentrer les tests sur un sous-ensemble d'échantillons pour lesquels les différences au niveau acoustique permettent de discriminer les systèmes évalués. Par ailleurs, on peut noter qu'aucune hypothèse n'a été faite sur la qualité de la sortie des systèmes.

Pour compléter l'évaluation de la méthode, nous avons vérifié sur les systèmes par corpus que

Système préféré	Corpus de coûts min.	Corpus aléatoire	Corpus de coûts max.
TTSCouv	27	34	52
CompAléa	27	37	32
Indifférent	46	29	16
Différence significative	Non	Non	**Oui**

(a) Résultats pour les systèmes par sélection d'unités. Trois tests AB ont été réalisées en présentant les échantillons les plus similaires, des échantillons aléatoires et les échantillons les plus différents.

Système préféré	Corpus aléatoire	Corpus de coûts max.
HMM-p3	31	26
HMM-p5	41	51
Indifférent	28	23
Différence significative	Non	**Oui**

(b) Résultats pour les systèmes à base de HMM. Deux tests AB ont été réalisées en présentant des échantillons aléatoires et les échantillons les plus différents.

TABLE 3: Résultats des tests de préférences

les échantillons les plus similaires donnent des résultats cohérents dans les tests perceptifs. Dans le tableau 3a, nous pouvons observer que, dans ce cas, un grand nombre d'échantillons sont jugés équivalents (46 votes *Indifférent*). Le reste des votes est réparti également entre *TTSCouv* et *CompAléa*. Encore une fois, la mesure appliquée ne donne aucune indication sur la qualité des échantillons. Pour conclure, ces résultats montrent clairement que sélectionner soigneusement les échantillons utilisés lors des tests perceptifs est primordial pour l'obtention de résultats significatifs.

6 Conclusion

Dans cet article, nous avons présenté une nouvelle méthode d'évaluation perceptive fondée sur un grand ensemble de test (des milliers d'échantillons) et une mesure utilisée pour classer les échantillons appariés en termes de différences acoustiques. Nous suggérons que les échantillons choisis doivent être les plus différents possibles afin d'être en mesure d'augmenter la significativité des évaluations perceptives. Cette nouvelle idée a été appliquée avec succès sur deux systèmes à base de HMM et deux systèmes par sélection d'unités, avec des voix d'apprentissage différentes (expressive avec locuteur masculin et neutre avec une locutrice). Les évaluations perceptives ont été menées pour comparer la méthode de sélection aléatoire classique à celle que nous proposons. Les résultats montrent clairement une amélioration de la significativité des résultats et une diminution des réponses de type « indifférent ».

Cette nouvelle méthodologie, qui reste simple, peut alors aider à valider efficacement des améliorations d'un système de synthèse vocale. Elle peut également être utilisée dans un procédé industriel en vue d'organiser des tests de non-régression entre les différentes versions d'un système avec un très faible coût et de repérer les phrases les plus touchées.

À l'heure actuelle, la méthode a été appliquée à des paires de systèmes, et les travaux futurs seront faits pour étendre cette méthode à un plus grand nombre de systèmes. Une approche pourrait être de réaliser les comparaisons par paires de systèmes puis de définir un rang moyen pour sélectionner les plus différents globalement. Nous prévoyons également de comparer la DTW avec d'autres distances de signaux acoustiques qui pourraient être mieux corrélées avec les évaluations perceptives.

Références

BOEFFARD O., CHARONNAT L., MAGUER S. L. & LOLIVE D. (2012). Towards fully automatic annotation of audio books for tts. In *Proc. of LREC*.

BUCHHOLZ S., LATORRE J. & YANAGISAWA K. (2013). Crowdsourced assessment of speech synthesis. *Crowdsourcing for Speech Processing*.

CHEVELU J., BARBOT N., BOËFFARD O. & DELHAY A. (2008). Comparing set-covering strategies for optimal corpus design. In *Proc. of LREC*.

CONKIE A., BEUTNAGEL M. C., SYRDAL A. K. & BROWN P. E. (2000). Preselection of candidate units in a unit selection-based text-to-speech synthesis system. In *Proc. of ICSLP*.

FRANÇOIS H. & BOEFFARD O. (2002). The greedy algorithm and its application to the construction of a continuous speech database. In *Proc. of LREC*.

GARCIA M.-N., D'ALESSANDRO C., BAILLY G., BOULA DE MAREÜIL P. & MOREL M. (2006). A joint prosody evaluation of french text-to-speech synthesis systems. In *Proc. of LREC*.

GUENNEC D. & LOLIVE D. (2014). Unit Selection Cost Function Exploration Using an A* based Text-to-Speech System. In *Proc. of TSD*.

HINTERLEITNER F., NEITZEL G., MOLLER S. & NORRENBROCK C. (2011). An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks. In *Proc. of Blizzard Challenge Workshop*.

ITU-T (1996). Recommendation : Methods for subjective determination of transmission quality.

KING S. & KARAISKOS V. (2012). The blizzard challenge 2012. In *Proc. of Blizzard Challenge workshop 2012*.

KRUL A., DAMNATI G., YVON F., BOIDIN C. & MOUDENC T. (2007). Adaptive database reduction for domain specific speech synthesis. In *Proc. of SSW6*.

LAMBERT T., BRAUNSCHWEILER N. & BUCHHOLZ S. (2007). How (not) to select your voice corpus : Random selection vs. phonologically balanced. In *Proc. of SSW6*.

LATORRE J., YANAGISAWA K., WAN V., KOLLURU B. & GALES M. J. (2014). Speech intonation for tts : Study on evaluation methodology. In *Proc. of Interspeech*.

LE MAGUER S., BARBOT N., BOËFFARD O. *et al.* (2013). Evaluation of contextual descriptors for hmm-based speech synthesis in french. In *SSW8*.

PRAHALLAD K., VADAPALLI A., KESIRAJU S., MURTHY H. A., LATA S., NAGARAJAN T., PRASANNA M., PATIL H., SAO A. K., KING S., BLACK A. W. & TOKUDA K. (2014). The blizzard challenge 2014. In *Proc. of Blizzard Challenge workshop 2014*.

SAINZ I., NAVAS E., HERNAEZ I., BONAFONTE A. & CAMPILLO F. (2014). Tts evaluation campaign with a common spanish database. In *Proc. of LREC*.

SAKOE H. & CHIBA S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*.

TOKUDA K., ZEN H. & BLACK A. W. (2002). An HMM-based speech synthesis system applied to English. In *IEEE Workshop 2002*.

VISWANATHAN M. & VISWANATHAN M. (2005). Measuring speech quality for text-to-speech systems : development and assessment of a modified mean opinion score (mos) scale. *Computer Speech & Language*.

ZEN H., TOKUDA K. & BLACK A. (2009). Statistical parametric speech synthesis. *Speech Communication*.

Constituance et phrasé prosodique en français : une étude perceptive.

Laury Garnier [1, 2] Corine Astésano [1] Lorraine Baqué [3] Anne Dagnac [2]
(1) URI Octogone-Lordat (E.A 4156), Université Toulouse 2 Jean-Jaurès, France
(2) Laboratoire CLLE-ERSS (UMR 5263), Université Toulouse 2 Jean-Jaurès, France
(3) Laboratori fLexSem, Universitat Autònoma de Barcelona, Espagne
laury.garnier@univ-tlse2.fr, astesano@univ-tlse2.fr,
lorraine.baque@uab.cat, dagnac@univ-tlse2.fr

RESUME

L'objectif de cette étude est d'explorer l'organisation du phrasé prosodique en français. Il n'existe pas de consensus clair sur le nombre de niveaux nécessaires pour refléter la hiérarchie prosodique de la langue. Dans ce cadre, nous proposons une étude perceptive, via un corpus de parole contrôlée manipulant des structures syntaxiques ambiguës, où 27 participants ont effectué 3 tâches de perception : proéminence, frontière et groupement. Nos résultats montrent une utilisation privilégiée des indices de frontières dans le marquage des groupes prosodiques. Plus précisément, on observe que les auditeurs sont capables de percevoir des niveaux de granularité de frontières plus fins que ce que les descriptions traditionnelles du français prédisent. Par ailleurs, les résultats de la tâche de proéminence montrent que l'accent initial est toujours perçu plus fort que l'accent final, et ce dès les niveaux les plus bas de la hiérarchie.

ABSTRACT

Prosodic constituency and phrasing in French: a perception study
The aim of the present study is to investigate the organization of prosodic phrasing in French. There is no clear consensus on how many levels are necessary to reflect the prosodic hierarchy in this language. In this context, we propose a perception study on a corpus manipulating syntactically ambiguous structures, where 27 participants had to perform 3 distinct perceptual tasks: prominence, boundary and grouping tasks. Our results show a preferential use of boundary cues in prosodic groups' marking. More precisely, we observe that listeners are able to distinguish finer-grained grouping levels than those predicted in traditional French descriptions. Moreover, the results of the prominence task show that initial accents are always perceived stronger than final accents, even at the lowest levels of the prosodic hierarchy.

MOTS-CLES : phrasé prosodique, perception, proéminence, frontière, groupement, français.
KEYWORDS: prosodic phrasing, perception, prominence, boundary, grouping, French.

1 Introduction

Les indices prosodiques, tels que les proéminences et les frontières, vont segmenter le flux de parole en groupes de mots pour ainsi faciliter la compréhension du message ; c'est ce qu'on appelle le phrasé prosodique. Là intervient alors le lien entre la prosodie et d'autres niveaux linguistiques,

notamment le niveau syntaxique. Un certain nombre de travaux se sont intéressés au lien à établir entre la structure prosodique et la structure syntaxique. Au-delà de la question du mapping entre les deux domaines, il apparaît difficile de trouver un consensus sur le choix des niveaux qui reflètent la hiérarchie prosodique. En effet, différentes hiérarchies prosodiques ont été proposées dans la littérature, variant de 2 niveaux (*syntagme accentuel* (ci-après *ap*) et *syntagme intonatif* (ci-après *IP*): Pierrehumbert, 1980, et Jun & Fougeron, 2002 pour le français) à 5 niveaux (incluant, selon les modèles : *mot prosodique* (ci-après *pw*), *groupe clitique, syntagme phonologique* (ci-après *PP*), *syntagme intonatif* et *énoncé* ; Selkirk, 1986 ; Nespor & Vogel, 1986 ; Post, 2000, notamment). Certains auteurs envisagent également la possibilité d'unités prosodiques récursives, dans le but de mieux rendre compte du lien entre prosodie et syntaxe. Parmi ces propositions, on retrouve le *« Super-Major Phrase »* (*super-syntagme majeur* : une *IP* pouvant contenir elle-même plusieurs *IP*) servant de niveau intermédiaire entre l'*énoncé* et le *syntagme majeur* (Ladd, 1996). De la même manière, Selkirk & Tateishi (1988) proposent de diviser le niveau du *PP* en 2 niveaux distincts : le *PP mineur* (équivalent à l'*ap*) et le *PP majeur* (équivalent au *syntagme intermédiaire* (ci-après *ip*) proposé par Beckman & Pierrehumbert, 1986). Plus récemment, ce principe de récursivité a été observé via la révision de la *Strict Layer Hypothesis* dans le cadre de la *Théorie de l'Optimalité* (Selkirk, 1995 ; Kager, 1999). L'existence d'un niveau intermédiaire a également été discutée pour le français. Quand certains auteurs considèrent l'*ip* comme une unité propre à des structures syntaxiques spécifiques, telles que les *« question tags »* ou les structures disloquées (Jun & Fougeron, 2000 ; voir aussi le *segment d'unité intonative* de Di Cristo & Hirst, 1996), d'autres, en revanche, proposent l'existence d'un « vrai » niveau intermédiaire entre le niveau de l'*ap* et de l'*IP*, qui permet de rendre compte non plus d'un type de structure syntaxique particulier mais davantage d'une unité de niveau supérieur à *l'ap,* fonction de la taille des constituants (Michelas & d'Imperio, 2010, en lien avec les propositions de Beckman & Pierrehumbert, 1986).

Pour le français, cette question autour des niveaux hiérarchiques est particulièrement intéressante. Alors que dans les contours intonatifs de l'anglais les proéminences se distinguent clairement des mouvements tonals associés aux frontières, le français, au contraire, est caractérisé par un syncrétisme entre les tons de frontière et les proéminences finales (ci-après *AF*), tous deux se manifestant sur la dernière syllabe des constituants prosodiques. Également, à l'inverse d'autres langues, l'accent en français est dit post-lexical et marquerait de manière privilégiée le niveau de l'*ap* plutôt que le niveau du *pw*[1]. Ces caractéristiques du français ont d'ailleurs poussé certains auteurs à parler de *langue sans accent* (Rossi, 1980) ou de *langue de frontière* (Vaissière, 1991). Plus récemment, dans le cadre de l'approche métrique autosegmentale, *AF* en français est décrit comme un *accent mélodique* (*pitch accent* : *H**), marquant le plus petit niveau de la hiérarchie prévu en français (*ap*). Dans cette conception, *AF* s'effacerait au profit du seul contour intonatif (*boundary tone* : *H%*) à un haut niveau de la hiérarchie prosodique : l'*IP* (Jun & Fougeron, 2002). Il existe également un autre type d'accent en français marquant le bord gauche de l'*ap* : il s'agit de l'accent initial (ci-après *AI*). Cet accent est dit secondaire et optionnel, *AF* étant primaire. Traditionnellement, *AI* est décrit comme un marqueur rythmique dont l'apparition dépend de la taille des constituants (Vaissière, 1991 ; Rossi, 1985). Toutefois, la description du rôle et des fonctions de cet accent dans le marquage même de la hiérarchie prosodique reste encore aujourd'hui mal établie. En effet, Astésano *et al.* (2007) se sont intéressés aux fonctions de *AI* via l'analyse acoustique d'un corpus contrôlé, manipulant à la fois la taille des constituants et la structure syntaxique. Dans les résultats de cette étude, *AI* se révèle être un marqueur plus structurel que purement rythmique ; les auteurs montrent que *AI* marque la structure prosodique de manière privilégiée par rapport à *AF*,

[1] *pw* est parfois compris comme l'équivalent du *groupe clitique* de Nespor & Vogel (1986), donc similaire à l'*ap*. Nous entendons ici *pw* dans l'acception de Selkirk (1996) qui correspond au *mot lexical*, les clitiques étant rattachés au niveau immédiatement supérieur.

spécifiquement à des niveaux inférieurs de la hiérarchie prosodique : le niveau de l'*ap*, voire le niveau du *pw*. Également, une étude en perception sur le même corpus confirme ces résultats (Astésano *et al.,* 2012). Dans cette dernière, les auteurs montrent que les auditeurs francophones sont capables de percevoir les proéminences finales (*AF*) indépendamment des frontières intonatives, et que *AI* est perçu systématiquement plus fort que *AF* à tous les niveaux de la hiérarchie prosodique.

Dans le prolongement de ces précédents travaux (Astésano *et al.* 2007 et 2012), nous proposons une étude perceptive, plus spécifiquement destinée à aborder la question des niveaux hiérarchiques de la structure prosodique en français. À travers la perception des francophones natifs, nous nous intéressons à la manière dont les indices de proéminences et de frontières sont utilisés pour instancier la structure morphosyntaxique. Plus précisément, cette étude vise à éprouver la hiérarchie à 3 niveaux (*ap*, *ip*, *IP*) proposée pour le français, sur la base de structures syntaxiques très contrôlées, afin de tester si ce fonctionnement à 3 niveaux est suffisant pour décrire le phrasé prosodique en français. Par ailleurs, la perception est utilisée ici comme une interface entre le niveau acoustique et le niveau phonologique permettant de mettre en évidence les indices pertinents dans le traitement du phrasé prosodique (comme proposé par Di Cristo, 2004 ; Cole et *al.*, 2010). Enfin, le prisme de la perception nous permet de contourner le problème de la variabilité ou *« flexibilité »* des indices acoustiques instanciant les constituants de même niveau dans la hiérarchie prosodique (Rossi, 1997 et 1999 ; Delais-Roussarie & Feldhausen, 2014).

2 Matériel linguistique et procédure expérimentale

Le matériel linguistique utilisé pour cette étude de perception est issu du *Corpus d'Edimbourg,* à la base de l'étude d'Astésano *et al.* (2007). Le corpus se constitue de structures syntaxiquement ambiguës, composées de deux noms coordonnés (*N1* et *N2*) et d'un adjectif (*A*), que les indices prosodiques (frontières et proéminences) aident à désambiguïser. Les deux structures syntaxiques possibles sont créées en manipulant la portée de l'adjectif : alors qu'en Condition 1, l'adjectif qualifie uniquement *N2* [*les lumières*] [*et les balises vertes*], en Condition 2, l'adjectif qualifie à la fois *N1* et *N2* [*les lumières et les balises*] [*vertes*]. Ces séquences *N1+N2+A* ont été placées dans un contexte phrastique plus large forçant ainsi la réalisation d'une frontière prosodique forte (frontière d'*IP*) après le syntagme cible. Selon la condition syntaxique, une structure prosodique différente est prédite, comprenant 3 niveaux au-dessous du niveau de l'*IP* : *ip*, *ap* et *pw*.

- **Condition 1 :** $[\{ (\textit{les} / \textit{lumières} /_{pw})_{ap} \}_{ip} \{ (\textit{et les} / \textit{balises} /_{pw} / \textit{vertes} /_{pw})_{ap} \}_{ip}]_{IP}$
- **Condition 2 :** $[\{ (\textit{les} / \textit{lumières} /_{pw})_{ap} (\textit{et les} / \textit{balises} /_{pw})_{ap} \}_{ip} \{ (/ \textit{vertes} /_{pw})_{ap} \}_{ip}]_{IP}$

Alors que le niveau de l'*ap* est communément accepté pour le français, celui de l'*ip* est controversé et le *pw* n'est même presque jamais mentionné. Cependant, la question de la pertinence de ce dernier niveau d'analyse avait déjà été soulevée dans les études acoustiques et perceptives précédentes d'Astésano *et al*. (2007 et 2012). La présente étude vise à approfondir l'investigation de cette granularité plus fine de la hiérarchie prosodique en français. Le corpus manipule également la taille des constituants (mots de 1 à 4 syllabes) et se compose de 4 sets de séquences *N1+N2+A*, lues par 8 locuteurs différents. Un test de jugement sémantique a été réalisé sur toutes les structures afin de ne garder que les locuteurs parvenant à instancier les deux conditions syntaxiques (cf. Astésano *et al.,* 2007 pour plus de détails sur la constitution du corpus). Pour cette étude perceptive, un seul des 4 sets lu par 1 locutrice a été utilisé : 32 syntagmes (4 longueurs de *N* * 4 longueurs de *A* * 2 conditions syntaxiques). Toutefois, nous avons exclu de notre analyse de données les séquences composées de *N* monosyllabiques puisqu'elles ne permettaient pas de distinguer *AI* et *AF*. Nos

résultats concernent donc 24 structures syntaxiquement ambiguës, soit 12 séquences par condition. Chacun de ces syntagmes a été jugé perceptivement par 27 auditeurs francophones natifs. Les auditeurs effectuaient 3 tâches de perception (cf. Figure 1) : une *tâche de proéminence* où ils devaient juger le degré de mise en relief de chaque syllabe du syntagme (variant de 6 à 15 sites potentiels selon la combinaison du nombre de syllabes des constituants), sur une échelle d'évaluation allant de 0 à 3 ; une *tâche de frontière* où les auditeurs devaient juger le degré de rupture entre chaque mot (5 sites potentiels) sur une échelle d'évaluation allant de 0 à 3 ; une *tâche de groupement* où ils devaient juger la manière dont les mots se regroupaient les uns avec les autres, parmi 4 propositions de groupements.

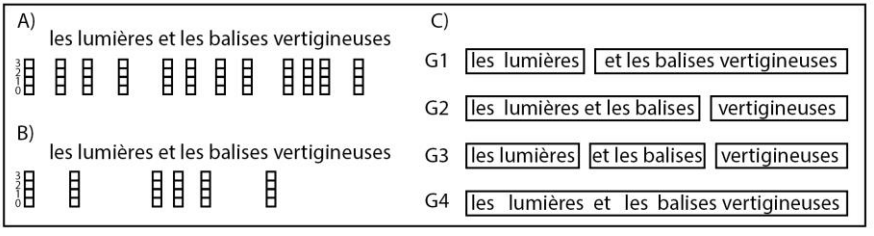


FIGURE 1 : A) Tâche de proéminence ; B) Tâche de frontière ; C) Tâche de groupement.

Chaque participant était muni d'un casque audio, confortablement assis devant un ordinateur, et pouvait écouter jusqu'à 5 fois le même syntagme. Avant chaque tâche, les participants effectuaient une phase d'entraînement, comprenant 4 séquences issues d'un set différent de celui utilisé dans cette étude. L'ordre de présentation des 3 tâches était contrebalancé entre tous les participants.

Notre étude de perception vise à préciser le rôle des indices prosodiques dans le marquage des unités morphosyntaxiques. Dans les deux premières tâches, proéminences et frontières sont volontairement jugées séparément, afin de rendre compte du rôle distinct de ces deux événements phonologiques dans le marquage de la structure. La troisième tâche de groupement sert plus spécifiquement à questionner les stratégies du phrasé prosodique, mais aussi à mettre en évidence les facteurs sous-jacents expliquant, notamment, les variations de correspondance potentielles avec la syntaxe. Alors que *G1* et *G2* correspondent aux deux conditions syntaxiques de départ (respectivement Condition 1 et Condition 2), *G3* et *G4* ont été choisis pour proposer une alternative à ces deux groupements, en permettant de faire ressortir des degrés de granularité de niveaux de frontières plus ou moins fins. *G3* satisfait à la fois les deux prédictions syntaxiques, bien qu'on attende néanmoins différentes forces relatives de frontières en Condition 1 (*N1* || *N2* | *A*) et en Condition 2 (*N1* | *N2* || *A*). *G4* peut également refléter les deux conditions syntaxiques, mais ne permet pas de mettre en évidence des frontières internes graduées à l'intérieur de *l'IP*.

L'analyse des scores de perception issus de ces 3 tâches suit 2 grandes étapes : une première phase consiste à tester l'adéquation des prédictions associées aux deux conditions syntaxiques avec les résultats de perception des groupements, en observant la manière dont les deux conditions sont distribuées à travers *G1*, *G2*, *G3* et *G4*. Dans une seconde étape, les résultats de groupements sont mis en relation avec les scores de frontières et de proéminences séparément, afin de faire émerger l'implication respective, mais aussi conjointe, de ces phénomènes prosodiques dans la structuration de la parole. Pour cette seconde analyse, la taille des constituants est prise en compte uniquement pour les noms, alors que pour les adjectifs elles sont regroupées afin d'augmenter le pouvoir statistique. Nous choisissons d'effectuer des analyses sur deux sites particuliers, points clés de la structure (cf. Figure 2) : entre *N1* et *N2* (*N1*|*N2*) et entre *N2* et *A* (*N2*|*A*). Pour chacun de ces sites,

nous relevons le score de frontière perçue entre chaque mot, ainsi que les scores de proéminences perçues sur les syllabes positionnées directement avant (*AF* potentiels sur *N1* et *N2*) et après (*AI* potentiels sur *N2* et *A*) la frontière (hors mots grammaticaux).

FIGURE 2 : Sites d'intérêt : *N1|N2* et *N2|A*.

Sur la base de ces données, nous construisons un modèle linéaire mixte avec le score de groupement et le nombre de syllabes du nom comme prédicteurs, et les participants et les items lexicaux comme variables aléatoires. Nous effectuons une analyse séparée pour chaque variable dépendante : le score de proéminence d'une part, et le score de frontière d'autre part. Les variables ou interactions non-significatives sont progressivement exclues du modèle final.

3 Résultats et discussions

3.1 Effet de la condition syntaxique sur la perception du groupement

Un test χ^2 d'indépendance montre un effet significatif de la condition syntaxique sur le groupement ($\chi^2(3,648)=530.89$, $p<.001$). Les analyses post-hoc indiquent que la Condition 1 est fortement associée à *G1* (93.5% *vs. G2* : 0.6%, *G3* : 5.9% et *G4* : 0.0% ; valeur de *p* ajustée <.001), alors que la Condition 2 est plus fréquemment associée à *G4* (42.9%), *G2* (32.1%) et *G3* (21.0%) qu'à *G1* (4.0% ; valeur de *p* ajustée <.001). A l'issue de ce test, nous décidons d'exclure de nos futures analyses les cas trop marginaux où la Condition 1 a été perçue comme *G2*, *G3* et *G4*, ainsi que les cas où la Condition 2 a été perçue comme *G1*. Nos résultats indiquent que les auditeurs sont capables de distinguer les deux conditions syntaxiques comme prédit : *G1* n'est quasiment pas associé à la Condition 2 et *G2* n'est pas perçu pour la Condition 1. En revanche, alors qu'on pouvait s'attendre à une répartition relativement équivalente de *G3* et *G4* sur les deux conditions, nous observons que ces deux groupements sont massivement associés à la Condition 2. Différents degrés de granularité dans le phrasé prosodique ont donc été perçus dans le cadre de cette dernière condition. La mise en relation de la perception des groupements avec les scores de frontières et de proéminences nous permet d'éclairer ce premier résultat (cf. infra).

3.2 Effets du groupement et de la longueur des *N* sur les scores de frontières

A) Force des frontières entre *N1|N2* (cf. Figure 3-A) :

Les résultats montrent un effet significatif du groupement ($F(3,91.25)= 494.14$, $p<.001$) et du nombre de syllabes du nom ($F(2,37.54)= 10.59$, $p<.001$), ainsi qu'une interaction entre ces deux prédicteurs ($F(6,89.32)= 4.04$, $p<.001$). Les tests post-hoc montrent des différences significatives ($p<.05$) des scores de frontières entre les différents types de groupements, quel que soit le nombre de syllabes du nom : la frontière en *G1* (étendue du score moyen (*sm*) : 3.38–3.62) est perçue beaucoup plus forte que dans les autres groupements (étendue du *sm* pour *G2* : 1.44–1.83 ; *G3* : 1.43–2.21 ; *G4* : 1.36–1.60). Également, la frontière en *G3* pour les noms de 2 et 4 syllabes est perçue plus forte

(*sm* respectifs : 2.10 et 2.21) qu'en *G2* (*sm* respectifs : 1.58 et 1.83) et *G4* (*sm* respectifs : 1.40 et 1.60). Cependant, pour les noms de 3 syllabes, nous n'observons pas de différences significatives (*sm* pour *G3* : 1.43 ; *G2* : 1.44 ; *G4* : 1.36). À l'issue de cette analyse, il semble donc que les auditeurs perçoivent 3 niveaux de frontière différents entre *N1|N2* : la frontière la plus forte en *G1*, intermédiaire en *G3*, et la plus faible en *G2* et *G4*.

B) Force des frontières entre *N2|A* (cf. Figure 3-B):

Nous observons un effet significatif du groupement ($F(3,111.64) = 30.95$, $p<.001$) seulement. Les tests post-hoc montrent une différence significative ($p<.05$) entre tous les groupements, excepté entre *G2* et *G3*. La frontière en *G1* (*sm* : 1.77) est perçue plus faible qu'en *G4* (*sm* : 1.99), qui, elle-même, est perçue plus faible qu'en *G2* (*sm* : 2.42) et *G3* (*sm* : 2.52). Les auditeurs semblent donc avoir perçu 3 niveaux de frontières différents entre *N2|A* : la frontière la plus forte en *G2* et *G3*, intermédiaire en *G4*, et la plus faible en *G1*.

Alors que la taille des constituants n'a pas d'effet sur le score des frontières, le type de groupement, en revanche, explique le degré de force de ces frontières. Néanmoins, ces analyses par sites (*N1|N2* et *N2|A*) ne permettent pas de rendre compte du rapport hiérarchique entre les frontières des deux sites. Nous proposons alors une analyse syntagmatique de la force relative des frontières en soustrayant le score de frontière perçu entre *N1|N2* à celui perçu entre *N2|A*.

C) Force relative des frontières entre les 2 sites (*N2|A - N1|N2*) (cf. Figure 3-C):

Les résultats montrent un effet significatif du groupement ($F(3,106.96) = 278.10$, $p<.001$) seulement. Les tests post-hoc indiquent que les seuls contrastes significatifs ($p<.001$) sont ceux entre *G1* (*sm* : -1.78) et les 3 autres groupements (*sm* pour *G2* : 0.96 ; *G3* : 0.47 ; *G4* : 0.50). En effet, pour *G1* seulement, la frontière sur le second site (*N2|A*) est perçue plus faible que celle sur le premier site (*N1|N2*). A l'inverse, pour *G2*, *G3* et *G4*, la frontière sur le second site (*N2|A*) est perçue plus forte que celle sur le premier site (*N1|N2*). Les résultats montrent donc que les auditeurs perçoivent deux types de hiérarchies prosodiques, avec une force relative des frontières différente, correspondant aux prédictions des deux conditions syntaxiques : rapport fort + faible pour la Condition 1 représentée ici par *G1* (*N1|N2 > N2|A*) et rapport faible + fort pour la Condition 2 associée ici à *G2*, *G3* et *G4* (*N1|N2 < N2|A*).

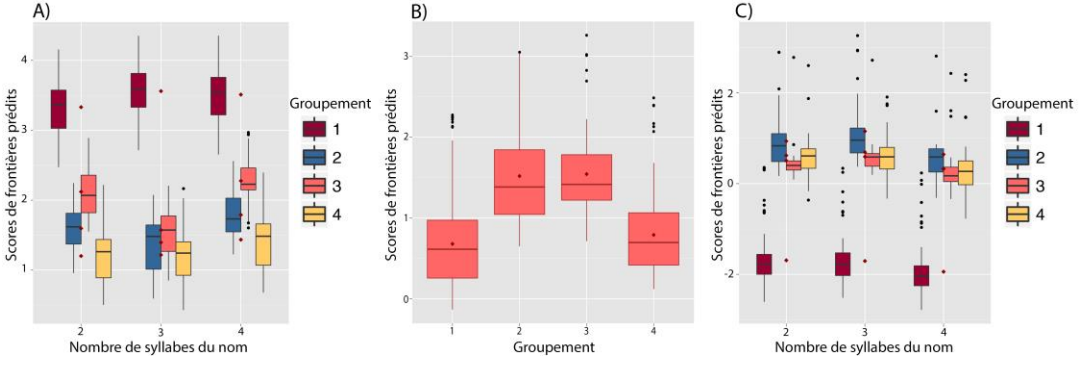


FIGURE 3: A) scores de la force des frontières entre *N1|N2* en fonction du groupement et du nombre de syllabes du nom ; B) scores de la force des frontières entre *N2|A* en fonction du groupement ; C) scores relatifs des frontières obtenus après soustraction des scores de frontières entre *N2|A* et *N1|N2*, en fonction du groupement et du nombre de syllabes du nom.

Nous observons une flexibilité dans la perception des frontières intra-site, expliquée par le type de groupement (cf. résultats A et B). Néanmoins, à travers l'analyse syntagmatique (cf. résultats C), on voit que la force relative des frontières inter-sites perçue par les auditeurs s'ajuste à la hiérarchie structurelle prédite par la syntaxe. Cette dernière analyse met également en évidence la nécessité de reconsidérer le nombre de niveaux hiérarchiques proposés pour la description du phrasé prosodique en français. En effet, on observe que selon la condition syntaxique, nous n'avons pas la même force relative entre les deux sites de frontières : rapport fort + faible en Condition 1 et rapport faible + fort en Condition 2. Nos résultats par sites nous permettent ensuite d'affiner l'interprétation de la granularité de ces 2 niveaux de frontières. En effet, il apparaît que les auditeurs perçoivent, sur chaque site, jusqu'à 3 niveaux de frontières inférieurs à l'*IP*. Entre *N1*et *N2*, la frontière perçue la plus forte en *G1* (***N1|N2*** **>** *N2|A*) pourrait correspondre à une frontière d'*ip*, alors que la frontière perçue un peu plus faible en *G3* (***N1|N2*** **<** *N2|A*) correspondrait à une frontière d'*ap*. Une frontière encore plus faible est perçue en *G2* et *G4* (***N1|N2*** **<** *N2|A*), impliquant potentiellement une frontière de *pw*. Entre *N2* et *A*, la frontière la plus forte perçue en *G2* et *G3* (*N1|N2* **<** ***N2|A***) pourrait correspondre à une frontière d'*ip*, la frontière perçue un peu plus faible en *G4* (*N1|N2* **<** ***N2|A***) correspondrait à une frontière d'*ap*. La frontière perçue encore plus faible en *G1* (*N1|N2* **>** ***N2|A***) impliquerait potentiellement, comme pour les résultats précédents, une frontière de *pw*. Ce résultat est particulièrement intéressant au regard des propositions faites sur les niveaux de constituance prosodique du français. Dans la plupart des approches théoriques, en effet, l'accent en français serait post-lexical, ce qui ne permet pas d'envisager un marquage de frontière à un niveau inférieur à l'*ap*, i.e. au niveau du *pw* proche du mot lexical chez Selkirk (1996). Or, il semble qu'il y ait bien un tel degré de niveau de frontière perçu par les auditeurs : entre *N2|A* en *G1* et entre *N1|N2* en *G2* et *G4*.

3.3 Effets du groupement et de la longueur des *N* sur les scores de proéminences

Pour les *AF* (sur *N1* et *N2*), nous notons un effet significatif du groupement ($F(3,568.71) = 6.24$, $p<.001$) seulement (effet trop marginal de la longueur des *N*) et ce, uniquement pour *AF* sur *N1*. Les tests post hoc indiquent que *AF* sur *N1* est perçu plus fort en *G1* (*sm* : 2.39, $p<.035$) qu'en *G2* et *G4* (*sm* respectifs : 2.09 et 2.17). Ce résultat montre donc qu'en *G1*, *AF* sur *N1* est clairement perçu sur le plan métrique, indépendamment de l'indice de frontière dans le marquage de la structure, et notamment ici dans le marquage de la frontière d'*ip*. Ce dernier point remet en question certaines propositions (Jun & Fougeron, 2002) selon lesquelles l'accent mélodique (*pitch accent* : *H**) disparaîtrait au profit du ton de frontière (*boundary tone* : *H%*) à un haut niveau de la hiérarchie prosodique. Concernant les *AI*, on n'observe ni d'effet du groupement, ni de la longueur des *N* ($p>.05$). Ce résultat questionne notamment le rôle purement rythmique de *AI* que les descriptions traditionnelles s'accordent à lui attribuer (Vaissière, 1991 ; Rossi, 1985). Pour aller plus loin, on propose alors de tester la contribution relative des proéminences initiales et finales dans le marquage des structures. Nous avons effectué un test de Student avec les scores d'*AI* et d'*AF* autour des deux sites *N1|N2* (*AF* sur *N1* et *AI* sur *N2*) et *N2|A* (*AF* sur *N2* et *AI* sur *A*), pour chaque groupement, sans tenir compte du nombre de syllabes du nom. Les résultats montrent une différence significative entre *AF* sur *N1* et *AI* sur *N2* : *AI* est perçu plus fort que *AF* en *G3* (*sm* : 1.63 *vs.* 1.07, $t(67) = 3.30$, $p=.002$) et en *G4* (*sm* : 1.56 *vs.* 1.07, $t(138) = 5.55$, $p<.001$). Également, *AI* est marginalement plus fort que *AF* en *G2* (*sm* : 1.34 *vs.* 1.11, $t(103) = 1.96$, $p=.053$). En revanche, il n'y a pas d'effet pour *G1* ($p=.338$). De la même manière, *AI* sur *A* est perçu significativement plus fort que *AF* sur *N2* pour tous les groupements ($p<.001$). Il semble donc que *AI* soit plus saillant perceptivement que *AF*, quel que soit le nombre de syllabes, et quel que soit le niveau dans la hiérarchie prosodique. Il est intéressant de noter que *AI* marque un niveau inférieur à l'*ap*. En effet, *AI* réalisé après la frontière du *pw* (entre *N2|A* en *G1*, entre *N1|N2* en *G2* et *G4*) est perçu plus fort que *AF* situé avant cette même frontière : *AI* sert donc ici à marquer la frontière gauche des *pw*.

4 Conclusion

L'objectif de ce travail était d'explorer, à travers une étude en perception, l'organisation du phrasé prosodique en français, et plus précisément l'utilisation des indices prosodiques dans le marquage de la structure. En effet, comme nous l'avons déjà exposé dans l'introduction, il n'existe pas de consensus clair sur le nombre de niveaux hiérarchiques nécessaires pour la description de la structure prosodique, et tout particulièrement pour la description du français. Deux résultats majeurs émergent : les participants utilisent de manière privilégiée les indices de frontières dans le marquage de la structure. Plus intéressant encore, les auditeurs sont capables de percevoir des niveaux de granularité de frontières plus fins que ce que les descriptions traditionnelles du français prédisent. En effet, nos résultats mettent en évidence la nécessité de prendre en compte un niveau supplémentaire dans la hiérarchie prosodique du français, inférieur à l'*ap* : le niveau du *pw*. Nos analyses sur les scores de proéminences vont également dans ce sens, en montrant que les *AI* sont plus saillants perceptivement que les *AF* et ce, dès le niveau du *pw*. Ces résultats, particulièrement intéressants sur la question des niveaux hiérarchiques nécessaires dans la structure prosodique du français, ouvrent la voie à de nouvelles analyses qui seront effectuées sur une plus large base de données, comprenant les mêmes syntagmes lus par 4 locuteurs différents (au total : 128 syntagmes) et entendus par 80 sujets. Avec un pouvoir statistique plus important, nous pourrons alors tester ces premiers résultats, en complétant notamment l'analyse des scores de proéminences avec d'autres sites accentuels (*AI* sur *N1 ; AF* sur *A*), mais également envisager des analyses acoustiques afin de comparer la réalité acoustique et perceptive des niveaux de constituance prosodique du français.

Remerciements

Cette étude a été réalisée dans le cadre du projet ANR-12-BSH2-0001 (IP: Corine Astésano) et soutenue par le ministère espagnol de l'économie et de la compétitivité FFI2013-40419-P (IP: Lorraine Baqué). Nous tenons également à remercier Rafèu Sichel-Bazin, avec qui nous avons pu partager de fructueuses discussions sur la constituance en français.

Références

ASTESANO C., BARD E.G., TURK A. (2007). Structural influences on initial accent placement in French. *Language and Speech*, 50(3), 423–446.

ASTESANO C., BERTRAND R., ESPESSER R., NGUYEN N. (2012). Perception des frontières et des proéminences en français. Actes des *JEP-TALN-RECITAL*, Grenoble, 353-360.

BECKMAN M.E., PIERREHUMBERT J.B. (1986). Intonational structure in Japanese and English. *Phonology*, 3(01), 255–309.

COLE J., MO Y., BAEK S. (2010). The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. *Lang. Cogn. Process.*, 25(7–9), 1141–1177.

DELAIS-ROUSSARIE E., FELDHAUSEN I. (2014). Variation in Prosodic Boundary Strength: a study on dislocated XPs in French. *Speech Prosody 2014*, Dublin, 1052–1056.

DI CRISTO A. (2004). La prosodie au carrefour de la phonétique, de la phonologie et de l'articulation formes-fonctions. *TIPA*, 23, 67–211.

DI CRISTO A., HIRST D. (1996). Rythme syllabique, rythme mélodique et représentation hiérarchique de la prosodie du français. *TIPA*, 15, 9–24.

JUN S-A., FOUGERON C. (2000). A Phonological Model of French Intonation. *Intonation: Analysis, Modeling and Technology*, Kluwer Academic Publishers, 15, 209–242.

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

KAGER R. (1999). *Optimality theory*. Cambridge : Cambridge University Press.

LADD D.R. (1996). *Intonational Phonology*. Cambridge : Cambridge University Press.

MICHELAS A., D'IMPERIO M. (2010). Durational cues and prosodic phrasing in French: evidence for the intermediate phrase. *Speech Prosody 2010*, Chicago, 100881:1-4.

NESPOR M., VOGEL I. (1986). *Prosodic phonology*. Foris: Dordrecht.

PIERREHUMBERT J.B. (1980). *The phonology and phonetics of English intonation*. Thesis, Massachusetts Institute of Technology, United States.

POST B. (2000). *Tonal and phrasal structures in French intonation*. Thesis, The Hague.

ROSSI M. (1980). Le français, langue sans accent? In *L'accent en français contemporain*, 15, I. Fonagy and P. Léon, Eds., 13–51.

ROSSI M. (1985). L'intonation et l'organisation de l'énoncé. *Phonetica*, 42 (2–3), 135–153.

ROSSI M. (1997). Is Syntactic Structure Prosodically Retrievable. *Proceedings of 5th European Conference of Speech Communication and Technology*, Greece, 1, 1–8.

ROSSI M. (1999). *L'intonation: le système du français : description et modélisation*. Paris, France: Editions OPHRYS.

SELKIRK E. (1986). On derived domains in sentence phonology. *Phonology*, 3, 371–405.

SELKIRK E. (1995). Sentence prosody: intonation, stress and phrasing. In *The Handbook of Phonological Theory*, Blackwell., J. Goldsmith, Ed. London, 550–569.

SELKIRK E. (1996). The prosodic structure of function words. In *Signal to syntax: Prosodic bootstrapping from speech to grammar in early acquisition*, J-L. Morgan and K. Demuth, Eds. Mahwah, NJ: Lawrence Erlbaum, pp. 187-214.

SELKIRK E., TATEISHI K. (1988). Constraints on Minor Phrase formation in Japanese. Proceedings of *Chicago Linguistic Society*, 24.

VAISSIÈRE J. (1991). Rhythm, accentuation and final lengthening in French. In *Music, Language, Speech and Brain*, Macmillan Press., J. Sundberg, L. Nord, and R. Carlson, Eds. London, 108–120.

Contribuer au progrès solidaire des recherches et de la documentation : la Collection Pangloss et la Collection AuCo

Alexis Michaud[1, 2] Séverine Guillaume[1] Guillaume Jacques[3] Đăng-Khoa Mạc[2] Michel Jacobson[4] Thu-Hà Phạm[5] Matthew Deo
(1) Langues et Civilisations à Tradition Orale, CNRS - Sorbonne Nouvelle - Institut national des langues et civilisations orientales, Paris, France
(2) Institut international de recherche MICA, Hanoi University of Science and Technology - CNRS - Grenoble INP, Hanoi, Vietnam
(3) Centre de Recherches Linguistiques sur l'Asie Orientale, CNRS - Ecole des Hautes Etudes en Sciences Sociales, Paris, France
(4) Laboratoire Ligérien de Linguistique, Universités d'Orléans et de Tours - BnF - CNRS
(5) Department of Linguistics, VNU University of Social Sciences and Humanities, Hanoi

alexis.michaud@vjf.cnrs.fr, severine.guillaume@vjf.cnrs.fr, rgyalrongskad@gmail.com, dang-khoa.mac@mica.edu.vn, michel.jacobson@gmail.com, phamha.ling@gmail.com, matthewdeo@gmail.com

RESUME

La présente communication présente les projets scientifiques et les réalisations de deux collections hébergées par la plateforme de ressources orales Cocoon : la Collection Pangloss, qui concerne principalement des langues de tradition orale (sans écriture), du monde entier ; et la Collection AuCo, dédiée aux langues du Vietnam et de pays voisins. L'objectif est un progrès solidaire des recherches et de la documentation linguistique. L'accent est mis sur les perspectives ouvertes pour la recherche en phonétique/phonologie par certaines réalisations récentes dans le cadre de ces deux Collections.

ABSTRACT

Contributing to joint progress in documentation and research: some achievements and future perspectives of the Pangloss Collection and the AuCo Collection

This talk sets out the scientific goals and achievements of two collections hosted by the Cocoon Open Archive of oral resources: the Pangloss Collection, which mainly focuses on unwritten languages from all areas in the world ; and the AuCo Collection, which is dedicated to languages of Vietnam and neighbouring countries. The aim is to contribute to joint progress in language documentation and in research. Emphasis is placed on the perspectives for phonetic/phonological research that are opened by some recent achievements in the framework of these two Collections.

MOTS-CLES : recherches phonétiques ; documentation linguistique ; archives orales ; archives ouvertes ; langues peu dotées ; diversité linguistique ; documentation en danger.

KEYWORDS: phonetic research; language documentation; language archives; open archives; under-resourced languages; linguistic diversity; endangered documentation.

1 Introduction

La présente communication présente les projets scientifiques et les réalisations de deux collections hébergées par la plateforme de ressources orales Cocoon : la Collection Pangloss, qui concerne principalement des langues de tradition orale (sans écriture), du monde entier ; et la Collection AuCo, dédiée aux langues du Vietnam et de pays voisins. L'objectif est un progrès solidaire des recherches et de la documentation linguistique. L'accent est mis sur les perspectives ouvertes pour la recherche en phonétique/phonologie par certaines réalisations récentes dans le cadre de ces deux Collections.

1.1 Etat des lieux : la prise de données à usage unique demeure pratique courante

Les bases empiriques des recherches phonétiques demeurent à l'heure actuelle un point de fragilité. La prise de données constitue un défi souvent sous-estimé (Niebuhr & Michaud 2015). Les bases de données sonores des centres de recherches en phonétique restent paradoxalement assez peu structurées, et relativement peu employées. Chercheurs et étudiants ont souvent tendance à constituer leur propre corpus en fonction des besoins de leur recherche, considérant qu'il est plus commode de recueillir de nouvelles données que de réemployer des ensembles documentaires existants. De fait, les recherches en phonétique nécessitent des données qui répondent à des critères précis concernant notamment les locuteurs et le type de méthodes expérimentales, critères que les données recueillies pour le propos d'expériences antérieures peuvent rarement satisfaire en intégralité. Les fonds d'archives restent relativement peu connus dans les laboratoires de phonétique. Les grands corpus distribués sur Internet peuvent être trop coûteux pour des recherches fondamentales (sans application commerciale directe) ; or les technologies numériques permettent d'enregistrer soi-même des données facilement et à faible coût. Des outils logiciels comme SpeechRecorder[1] permettent en outre de simplifier considérablement le travail d'édition et d'annotation des enregistrements.

On voudrait souligner ici les limites de cette logique : il est illusoire de penser que l'on peut à tout moment créer le corpus dont on a besoin. Dans le cas des langues en danger, la mise en commun des données existantes est particulièrement nécessaire, et des activistes de la documentation soulignent depuis des années l'importance de la conservation et la diffusion des données (voir notamment Thieberger et al. 2016). Mais dans l'étude des grandes langues, la prise de données gagnerait également à être conçue dans une logique de progrès cumulatif de la documentation, au lieu de collecter des données à usage unique (dont le réemploi n'est pas prévu d'emblée). Quiconque a l'expérience de la collecte de données confirmera qu'il s'agit d'une activité qui s'avère chronophage au final. Les étapes sont nombreuses : mise au point du protocole expérimental, tests, rendez-vous avec les participants, enregistrement, mise en forme... L'absence de réutilisation des données constitue une déperdition pour la communauté des recherches en parole.

Si l'on enregistre des données en ayant en tête la perspective d'un archivage pérenne et d'un partage auprès de la communauté, cela encourage à accroître l'investissement initial de temps et de soin, ce qui a, selon notre expérience, des conséquences positives en termes de qualité des données, et partant, de fiabilité des recherches. Pour ne prendre qu'un exemple, celui du choix des locuteurs : les usagers des laboratoires de phonétique sont souvent sollicités comme sujets pour enregistrer des

[1] http://www.bas.uni-muenchen.de/Bas/software/speechrecorder/

données. Ils ont l'expérience des tâches demandées, tandis que les non-initiés peuvent être intimidés ou perplexes ; par ailleurs, étudiants et collègues peuvent rendre service bénévolement. Mais le fait de recourir à un locuteur linguiste, souvent polyglotte, pose des problèmes épistémologiques évidents. Des mots français enregistrés par un locuteur natif pour un cours de lecture de spectrogrammes se sont avérés « non canoniques » au point d'induire en erreur des déchiffreurs chevronnés ; c'est vraisemblablement la conséquence d'une expérience linguistique diversifiée. Il n'y a rien de surprenant à ce que des locuteurs du japonais ou du vietnamien, après plusieurs mois en France, transposent dans leur langue maternelle les continuations intonatives qu'ils ont appris à employer en français (Dô Thê Dung, Trân Thien Huong & Boulakia 1998). Outre leur prononciation, ce séjour perturbe également leur façon de percevoir. Cela remet en cause certaines données publiées dans les revues internationales, fournies par des locuteurs natifs mais résidents depuis fort longtemps dans un pays étranger.

Ces fragilités restent actuellement masquées par le fait que les données ne sont généralement pas communiquées aux évaluateurs des travaux soumis aux revues scientifiques, ni aux lecteurs de ces travaux dans leur version publiée. En l'absence d'exigence de communication des données, que ce soit de la part des revues scientifiques ou des institutions qui financent le chercheur, pourquoi s'imposer des efforts supplémentaires, qui n'aideront pas à la publication des recherches, et ne compteront pas dans l'évaluation du chercheur ? Rendre ses données disponibles, c'est également prêter le flanc à la critique. Faudrait-il réécouter tous les enregistrements, pour vérifier qu'il n'y traîne pas un passage à écarter avant diffusion : fou-rires, raclements de gorge ou autres maladresses qui fourniraient matière à un montage audio tournant en ridicule les locuteurs et les auteurs ? La conclusion paraît claire : on a beaucoup à perdre (à commencer par un temps précieux) à partager ses données, et rien à y gagner.

Cette situation paradoxale n'a pas fondamentalement changé depuis un état des lieux présenté aux Journées d'Etude de la Parole il y a quatorze ans (Michaud 2002). On aimerait néanmoins essayer ici d'argumenter qu'on a beaucoup à gagner à introduire, dans sa pratique de recherche, un raisonnement en termes de progrès cumulatif de la documentation linguistique, solidaire de progrès de la recherche.

1.2 Problématique : associer documentation et recherche, pour leur bénéfice mutuel

Est-ce un hasard si les centres de recherche en phonétique qui mettent à disposition leurs collections sonores, sans être retenus par la crainte des critiques que pourraient attirer la qualité inégale des enregistrements, sont des pionniers mondiaux du domaine ? Le laboratoire de l'Université de Californie à Los Angeles a choisi de mettre en ligne des ressources abondantes, qui figurent en bonne place sur leur site internet (http://www.phonetics.ucla.edu/). Faut-il conclure que le temps consacré à l'archivage et la diffusion de données soit un luxe réservé aux chercheurs qui, forts d'un succès incontesté et d'une situation professionnelle assurée, peuvent se permettre des activités non rentables en termes de carrière ? Il nous paraît au contraire que le souci d'associer documentation et recherche est l'un des facteurs de la réussite du laboratoire de phonétique de UCLA. La valorisation du socle empirique de la recherche procède de la même logique qui a abouti au livre *The Sounds of the World's Languages* (Ladefoged & Maddieson 1996). L'une et l'autre de ces productions nous paraissent refléter la vision d'avenir de pionniers qui font le point des données et des analyses qu'ils ont dans leurs cartons, et qui publient cet inventaire de l'état de l'art, en le présentant pour ce qu'il est : une étape sur le chemin d'un progrès vers une compréhension plus complète de la face sonore des langues du monde.

2 La Collection Pangloss et ses usages pour la phonétique et le traitement automatique des langues

2.1 Présentation

La Collection Pangloss est une archive publique qui contient plus de 2.000 enregistrements (400 heures) en plus de 130 langues et dialectes, dont 990 documents annotés par une vingtaine de chercheurs. La Collection Pangloss réunit des documents linguistiques sonores, avec une spécialité de langues « rares » ou peu étudiées. Son but est de contribuer à la documentation et à l'étude du patrimoine humain que représentent les langues du monde.

La Collection Pangloss donne accès aux enregistrements sonores d'origine aussi bien qu'aux transcriptions et traductions ; c'est une garantie d'authenticité et une ressource pour la recherche. Les ressources associent donc son et texte. L'aspect texte comprend une transcription phonologique accompagnée, selon les cas, de représentations orthographiques (là ou celles-ci existent, y compris dans des écritures non latines), de traductions en diverses langues (généralement : anglais, français, et/ou langue nationale du pays d'enquête), de gloses morphologiques, de notes, etc. L'alignement de la transcription avec le son se fait généralement au niveau de la « phrase » ou du groupe intonatif, mais peut se faire également au niveau du mot ou du morphème.

La parole spontanée forme la plus grande partie du fonds : des documents enregistrés dans leur contexte social et transcrits en consultation avec les locuteurs. Mais la Collection contient également des séances d'enquête et des listes de mots, enregistrés et annotés par des chercheurs d'horizons très variés. La pérennité de ces ressources « rares » est assurée par Cocoon[2], archive structurée selon les normes actuelles (XML, OLAC, Dublin Core...), dans un format ouvert. Aussi bien ces données que les outils qui servent à leur préparation et leur diffusion sont librement disponibles sur le site de la Collection Pangloss[3]. Pour plus de détails concernant l'historique du projet, les choix technologiques et l'état actuel des collections, un article complet est disponible en ligne (Michailovsky et al. 2014) ; le présent exposé s'attache plus spécifiquement aux projets et collaborations possibles avec la communauté des phonéticiens/phonologues.

2.2 Usages pour l'enseignement et la recherche en phonétique

La Collection Pangloss se prête à divers usages pour l'enseignement et la recherche en phonétique ; usages qui, en retour, fournissent l'occasion d'un enrichissement.

Tout d'abord, on y trouve une illustration d'un grand nombre de sons des langues du monde. La langue oubykh, aujourd'hui éteinte, était l'une des deux langues les plus riches en consonnes jamais observées. Les enregistrements disponibles en ligne comportent diverses paires minimales enregistrées avec soin, qui ont servi à des tests de perception, et se prêtent aisément à une exploitation pour des enseignements en phonétique.

[2] *Collections de corpus oraux numériques* (Cocoon): http://cocoon.huma-num.fr Au sujet de cette plate-forme technique pour la gestion de collections de ressources orales, voir Jacobson et al. (2015), qui présente son fonctionnement ainsi que les innovations et expérimentations en cours, dont bénéficie l'ensemble des collections hébergées.

[3] http://lacito.vjf.cnrs.fr/pangloss/

Les données de la Collection Pangloss peuvent en outre fournir la matière à de nouvelles recherches au sujet de systèmes sonores. Pour reprendre l'exemple de l'oubykh, des enregistrements de cinéradiographie ont été réalisés, et ont donné lieu à une « Etude articulatoire de quelques sons de l'oubykh d'après film aux rayons X » (Leroy & Paris 1974). Les films aux rayons X ont été numérisés en 2001, mais n'ont pas encore été alignés avec les transcriptions, ni diffusés. Un travail de recherche au sujet de ces données, par exemple une étude comparée de certains types de consonnes dans diverses langues (pour le français: Bothorel et al. 1986), fournirait l'occasion de les préparer pour une mise en ligne, ce qui enrichirait l'ensemble documentaire.

Des applications en traitement automatique de la parole sont également envisageables. Les données de la Collection Pangloss comportent des annotations de grande qualité, fruit du travail de linguistes qui consacrent souvent la majeure partie de leur carrière à une langue ou un petit groupe de langues. Ces données, malgré leur faible volume, présentent par là un intérêt pour le traitement automatique : linguistique de corpus (concordances et étude des collocations), reconnaissance automatique de la parole, synthèse de la parole, ou traduction automatique.

A titre d'exemple : données de langue na de Yongning

Une communication soulignant la possibilité d'appliquer des traitements automatiques au corpus de langue na de Yongning (Michaud et al. 2012), assortie d'une étude-pilote en reconnaissance automatique (Do, Michaud & Castelli 2014), a attiré l'attention d'équipes travaillant en reconnaissance automatique, qui ont entrepris d'utiliser ces données de la Collection Pangloss pour des recherches d'avant-garde : tester les possibilités de parvenir à une transcription phonétique sans entraînement préalable. La transcription du linguiste sert de référence *(gold standard)* pour évaluer le degré de précision atteint par les algorithmes. On peut imaginer de nombreux autres scénarios, par exemple une collaboration entre chercheurs en informatique et en linguistique pour l'étude de la réalisation phonétique des séquences tonales. Un algorithme de reconnaissance automatique des tons serait entraîné au moyen du corpus déjà transcrit par le linguiste (environ cinq heures de parole). Un dialogue entre linguistes et informaticiens permettrait d'améliorer simultanément l'outil informatique et la modélisation linguistique, en testant l'emploi de paramètres proposés par le linguiste. Cela permettrait d'affiner les hypothèses au moyen d'une implémentation informatique. On pourrait par exemple imaginer de déterminer quelle proportion de l'information contenue dans la courbe de fréquence fondamentale peut s'expliquer par l'identité phonologique du ton (dans cette langue : Bas, Moyen, Haut, Bas-Haut, ou Moyen-Haut) ; par la déclinaison (schéma global d'évolution de la fréquence fondamentale au cours de l'énoncé) ; par le découpage en constituants ; et par la structure de l'information. Il pourrait être intéressant de montrer, par exemple, qu'en l'absence de prise en compte des paramètres contextuels (coarticulation tonale, déclinaison...), le pourcentage d'identification exacte n'est que de tant pour cent ; que la prise en compte de la coarticulation tonale améliore le système de tant pour cent ; que l'introduction d'une distinction entre mots pleins et mots-outils apporte une amélioration de tant, et ainsi de suite. A mesure des progrès de l'outil de reconnaissance développé pour la langue-cible, le résidu (tons non reconnus) serait de plus en plus limité, et de plus en plus intéressant pour une analyse qualitative : le linguiste pourrait exercer sa sagacité sur les passages qui induisent en erreur les algorithmes. L'expérience montrerait dans quelle mesure ce résidu tient à des questions purement techniques (par exemple la plus grande difficulté à traiter des syllabes réalisées avec une phonation non modale), et dans quelle mesure il soulève des questions intéressantes pour la modélisation.

2.3 Une nouvelle direction : les dictionnaires en ligne

La réalisation de dictionnaires en ligne, associés aux textes, fait partie du projet de la Collection Pangloss depuis ses débuts. Dans le cadre d'un projet coordonné par G. Jacques[4], trois nouveaux dictionnaires sont en ligne, librement consultables au format HTML ainsi que sous forme de documents PDF (composés en LaTeX). Le dictionnaire japhug-chinois-français comporte près de 8.000 entrées, ce qui en fait un modèle du genre. Une version consultable via smartphone est en cours de réalisation. Ces dictionnaires suivent la norme ISO LMF (Lexical Markup Framework), conçue pour permettre un traitement automatique. L'emploi du format-pivot XML permet des passerelles vers d'autres standards, tels que TEI (Text Encoding Initiative) (Romary 2013). Une lemmatisation systématique des textes permettrait d'accéder à partir du dictionnaire à toutes les occurrences du mot en contexte, et inversement, d'accéder à l'entrée de dictionnaire par un clic sur n'importe quel mot d'un texte. De nombreuses autres pistes sont imaginables pour des collaborations, qui permettraient de réaliser le fort potentiel de ces données.

3 La Collection AuCo

3.1 Présentation

La Collection AuCo est, comme la Collection Pangloss, un projet porté par un centre de recherche (l'Institut International de Recherche MICA, Unité Mixte Internationale CNRS-HUST-Grenoble INP, situé à Hanoi) mais qui a vocation à rendre service à une communauté plus étendue. La Collection regroupe des documents linguistiques sonores du Vietnam et des pays voisins. AuCo est un acronyme pour "Audio Corpora": corpus audio. C'est également une référence à la fée ÂuCơ, qui mit au monde une grande poche d'où sortirent cents œufs qui donnèrent naissance aux Cent Peuples, ancêtres légendaires de la multitude de groupes ethniques de la région. Les points qui composent le logo de la Collection AuCo sont une allusion à ces cent œufs, symbole de la diversité culturelle et linguistique que reflète la collection.

Le but de la Collection AuCo est de rassembler les documents recueillis par les chercheurs au fil de leur activité de recherche, contribuant ainsi à la documentation du patrimoine humain que représentent les langues. La Collection a aussi vocation à encourager et faciliter les travaux de recherche interdisciplinaires associant ingénieurs et linguistes, autour de techniques communes.

La collection accueille des documents de types très divers, et de valeur patrimoniale très inégale : des récits traditionnels aux documents lus, en passant par les dialogues et les enquêtes de vocabulaire ; des collections uniques datant de plusieurs décennies, et concernant des parlers aujourd'hui en voie de disparition, jusqu'au tout-venant des enregistrements de langues nationales (réalisés ponctuellement pour les besoins d'études phonétiques/phonologiques ou d'outils de traitement automatique). Les utilisations nouvelles et créatives des données sont rarement prévisibles; d'où le choix de ne fermer la Collection AuCo à aucun type de données.

[4] http://himalco.huma-num.fr/dictionaries/

3.2 Exemples d'études dont la base empirique est accessible librement

La Collection AuCo, tout comme la Collection Pangloss, offre aux chercheurs la possibilité d'archiver et rendre accessibles les données sur lesquelles reposent leurs travaux. A titre d'exemple, l'étude d'un dialecte de la langue vietnamienne, Phong Nha, a été entreprise en janvier 2014, et les résultats de l'enquête ont été publiés en 2015 (Michaud, Ferlus & Nguyễn 2015). L'enquête reposait sur une liste de vocabulaire, la liste « EFEO-CNRS-SOAS », disponible en ligne (Pain et al. 2014). Les fichiers audio enregistrés lors de l'enquête ont été annotés, en indiquant pour chaque item son numéro dans cette liste ; un script PRAAT (disponible en ligne) est appliqué aux fichiers TextGrid, pour produire des documents XML multilingues (français, anglais, vietnamien, chinois, khmer) synchronisés avec l'enregistrement. Les documents sont en ligne depuis fin 2015. Ainsi, les lecteurs intéressés de découvrir que la spirante /ð/ du vietnamien moyen est préservée dans le dialecte de Phong Nha (alors que dans le delta du Fleuve Rouge elle s'est confondue avec les consonnes /r/ et /ʒ/ ; toutes trois sont actuellement réalisés /z/) peuvent écouter de nombreux exemples, et se faire une opinion au sujet du périmètre de variation allophonique de ce son en fonction de la voyelle et du ton auxquels il se trouve associé.

Le travail de préparation des données a bien sûr demandé un investissement de temps, mais celui-ci a été réduit au minimum. On a fait l'économie du toilettage des fichiers audio : retrancher les portions de silence, d'apartés entre enquêteurs, de raclements de gorge... Les fichiers audio déposés en ligne ne sont pas prévus pour une écoute linéaire, mais pour un accès direct aux mots annotés, auxquels l'interface de consultation donne un accès direct par un bouton « lecture » associé à chacun des items. On pourrait aller jusqu'à argumenter que la mise en ligne des séances des deux journées d'enquête sans aucune retouche présente un avantage : ces documents illustrent le déroulement d'une enquête, et permettent de connaître le contexte de réalisation de chacun des mots. On pourrait par exemple imaginer de mesurer le temps de réponse : entre l'élicitation (le mot en vietnamien standard, fourni oralement) et la réponse du premier locuteur. Cette information pourrait un jour être utilisée dans le cadre d'une étude statistique des pratiques d'élicitation de vocabulaire.

3.3 Les collections de Michel Ferlus : données de plus de quarante parlers

De septembre 2014 à février 2016, dans le cadre de la Collection AuCo a été réalisé un projet de numérisation intitulé « DO-RE-MI-FA : Données des Recherches linguistiques de Michel Ferlus en Asie du sud-est ». Ce projet concerne l'ensemble des enregistrements audio réalisés par Michel Ferlus au cours de son activité comme « linguiste de terrain », de 1963 à 2003 : environ 200 heures d'enregistrements. Michel Ferlus est un spécialiste de la phonétique historique des langues d'Asie du Sud-Est (voir notamment Ferlus 1992, 1998). Ses données inédites proviennent de plus de 40 parlers, dont un grand nombre étaient jusque-là non documentés. Le projet bénéficiait d'une subvention de la Bibliothèque Scientifique Numérique du Ministère de l'Enseignement Supérieur et de la Recherche, dans le cadre de l'opération de numérisation du patrimoine de l'enseignement supérieur et de la recherche. L'objectif est de transformer le fonds de chercheur de Michel Ferlus en un ensemble documentaire dans les règles de l'art, pleinement exploitable, mis à la libre disposition de la communauté des chercheurs ainsi que d'un public plus étendu.

L'enrichissement de ces données par une annotation multilingue s'appuyant sur les notes de Michel Ferlus constitue une entreprise pour le moyen terme, indissociable de la formation de la jeune génération des chercheurs dans ce domaine scientifique. Cette démarche a été engagée dans le cadre du projet, avec la réalisation d'annotations pour la plupart des documents arem, mường, thổ (groupe

vietique de la famille austroasiatique), ainsi que pour une trentaine de documents de langues tai-kadai, par des personnes dont certaines contribueront à prendre la relève du travail de recherche de Michel Ferlus.

Pour le recollement entre enregistrements et transcriptions, le mode opérationnel actuel est le suivant : les manuscrits de Michel Ferlus sont saisis un par un, et leur contenu publié, soit par l'auteur lui-même, soit par un étudiant-chercheur (doctorant) intéressé à reprendre le flambeau pour l'étude d'une langue en particulier, et ayant une certaine familiarité avec la langue. La justification de ce choix est qu'un chercheur engagé dans l'étude de la langue-cible sait tirer le meilleur parti des notes de terrain, et redresser de lui-même les inévitables petites erreurs ou approximations dans la notation. Ce mode de fonctionnement peut paraître extrêmement contraignant : il se peut qu'il faille attendre plusieurs années avant que les données d'une certaine langue trouvent un utilisateur fort d'une bonne connaissance préalable du domaine linguistique concerné. Pour autant, cette perspective n'est nullement utopique : des collègues intéressés se manifestent plus fréquemment que l'équipe du projet ne l'avait initialement espéré.

Au plan technique, le développement et le déploiement d'un outil pour l'affichage synchronisé de manuscrits (en mode image) avec leur annotation (XML) et l'enregistrement audio du texte lu a été engagé dans le cadre du projet DO-RE-MI-FA. Ce logiciel, EASTLing (Easy Annotation & Synchronization Tool for Linguists), comprend un outil d'édition et un outil de lecture (voir : http://moon-light.fr/). Cette innovation illustre l'utilité de collaborations pour le partage de savoir-faire et l'enrichissement de l'interface de consultation des archives, solidaire d'un enrichissement des ressources.

4 Un mot de conclusion

Cette communication a atteint son but si elle est parvenue à mieux faire connaître des phonéticiens la Collection Pangloss et la Collection AuCo, et à suggérer le fort potentiel que présente une association renforcée entre documentation linguistique et recherche. Dans le détail, utilisations et nouveaux développements sont à inventer entre collègues intéressés : projets de recherche, projets applicatifs, ou encore développement d'interfaces « 2.0 » offrant aux utilisateurs le moyen de contribuer aux collections en signalant des corrections, en ajoutant des informations complémentaires, voire en déposant eux-mêmes de nouveaux documents.

Remerciements

Les réalisations présentées ici sont le produit du travail d'un grand nombre de personnes appartenant à divers corps de métier. On a respecté la consigne d'une publication nominale, mais un choix plus cohérent à nos yeux serait de signer du nom de « l'équipe Pangloss », « l'équipe AuCo » et « l'équipe Cocoon » réunies. Nous souhaitons remercier tout particulièrement Anne Behaghel ; Rémy Bonnet ; Céline Buret ; Eric Castelli ; Hằng Đinh ; Michel Ferlus ; Alexandre François ; Julien Heurdier ; Aimée Lahaussois ; Martine Mazaudon ; Boyd Michailovsky ; Minh-Châu Nguyễn ; Việt Sơn Nguyễn ; Thơ Nông ; Frédéric Pain ; Đỗ-Đạt Trần ; et Trí-Dõi Trần. Nous sommes vivement reconnaissants envers les institutions et structures partenaires suivantes : CNRS-InSHS ; Très Grande Infrastructure de Recherche *Humanités Numériques* (Huma-Num) ; CINES ; CC-IN2P3 ; ANR (projets *Empirical Foundations of Linguistics*, ANR-10-LABX-0083, et *HimalCo*, ANR-12-CORP-0006) ; Bibliothèque Scientifique Numérique (projet DO-RE-MI-FA, 2014-2016).

Références

BOTHOREL, André, Péla SIMON, François WIOLAND & Jean-Pierre ZERLING (1986). *Cinéradiographie des voyelles et des consonnes du français*. Strasbourg: Travaux de l'Institut de Phonétique de Strasbourg.

DO, Thê Dung, Thien Huong TRAN & Georges BOULAKIA (1998). Intonation in Vietnamese. In Daniel Hirst & Albert Di Cristo (eds.), *Intonation systems: a survey of twenty languages*, 395–416. Cambridge: Cambridge University Press.

DO, Thi-Ngoc-Diep, Alexis MICHAUD & Eric CASTELLI (2014). Towards the automatic processing of Yongning Na (Sino-Tibetan): developing a "light" acoustic model of the target language and testing "heavyweight" models from five national languages. *Proceedings of the 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2014)*, 153–160. St Petersburg. http://halshs.archives-ouvertes.fr/halshs-00980431/.

FERLUS, Michel (1992). Essai de phonétique historique du khmer (du milieu du premier millénaire de notre ère à l'époque actuelle). *Mon-Khmer Studies* 21. 57–89.

FERLUS, Michel (1998). Les systèmes de tons dans les langues viet-muong. *Diachronica* 15(1). 1–27.

JACOBSON, Michel, Nicolas LARROUSSE & Marion MASSOL (2015). La question de l'archivage des données de la recherche en SHS (Sciences Humaines et Sociales). *Archives et données de la recherche (ICA/SUV 2014)*. Paris. http://halshs.archives-ouvertes.fr/halshs-01025106.

LADEFOGED, Peter & Ian MADDIESON (1996). *The Sounds of the World's Languages*. Dirigé par M. Kenstowicz, J. Goldsmith, Nick Clements & D. Steriade. Oxford: Blackwell.

LEROY, Christine & Catherine PARIS (1974). Etude articulatoire de quelques sons de l'oubykh d'après film aux rayons X. *Bulletin de la Société de Linguistique de Paris* LXIX(1). 255–286.

MICHAILOVSKY, Boyd, Martine MAZAUDON, Alexis MICHAUD, Séverine GUILLAUME, Alexandre FRANÇOIS & Evangelia ADAMOU (2014). Documenting and researching endangered languages: the Pangloss Collection. *Language Documentation and Conservation* 8. 119–135.

MICHAUD, Alexis (2002). Conservation des langues et partage des ressources : le rôle des chercheurs dans la mise en place de banques de données. *XXIVe Journées d'Etude de la Parole*, 153–156. Nancy, France.

MICHAUD, Alexis, Michel FERLUS & Minh-Châu NGUYỄN (2015). Strata of standardization: the Phong Nha dialect of Vietnamese (Quảng Bình Province) in historical perspective. *Linguistics of the Tibeto-Burman Area* 38(1). 124–162. doi:10.1075/ltba.38.1.04mic.

MICHAUD, Alexis, Andrew HARDIE, Séverine GUILLAUME & Martine TODA (2012). Combining documentation and research: Ongoing work on an endangered language. In Xiong Deyi, Eric Castelli, Dong Minghui & Pham Thi Ngoc Yen, (eds.), *Proceedings of IALP 2012 (2012 International Conference on Asian Language Processing)*, 169–172. Hanoi, Vietnam: MICA Institute, Hanoi University of Science and Technology.

NIEBUHR, Oliver & Alexis MICHAUD (2015). Speech data acquisition: the underestimated challenge. *KALIPHO - Kieler Arbeiten zur Linguistik und Phonetik* 3. 1–42.

PAIN, Frédéric, Michel FERLUS, Alexis MICHAUD & Thu Hà PHẠM (2014). *EFEO-CNRS-SOAS word list for linguistic fieldwork in Southeast Asia*. Hanoi: International Research Institute MICA. https://halshs.archives-ouvertes.fr/halshs-01068533/.

ROMARY, Laurent (2013). TEI and LMF crosswalks. *arXiv preprint arXiv:1301.2444*.

THIEBERGER, Nick, Anna MARGETTS, Stephen MOREY & Simon MUSGRAVE (2016). Assessing annotated corpora as research output. *Australian Journal of Linguistics* 36(1). 1–21. doi:10.1080/07268602.2016.1109428.

Contribution à l'étude de la focalisation prosodique en français[1]

Rémi Godement-Berline
Univ. Paris Diderot, Sorbonne Paris Cité
LLF (UMR 7110), 8 place Paul Ricœur, 75013 Paris, France
remi.godement@linguist.univ-paris-diderot.fr

RESUME

Cette étude porte sur la focalisation prosodique en français dans plusieurs styles de parole (parole spontanée et lecture ou interprétation par des acteurs). Nous attribuons à la focalisation des fonctions sémantico-pragmatiques ou emphatiques. Un groupe de dix experts en prosodie a relevé les occurrences de focalisation dans le corpus d'étude. Les résultats confirment que la focalisation est réalisée par une augmentation de hauteur et de durée. Ils diffèrent de la littérature précédente du point de vue du type de contour prosodique employé sur les occurrences de focalisation et de la présence d'accent initial. Des problèmes méthodologiques sont soulevés concernant l'analyse des contours terminaux et de la désaccentuation.

ABSTRACT

Contribution to the study of prosodic highlighting in French.
This paper studies prosodic highlighting in French in several types of speech (spontaneous speech and reading aloud or performing by actors). We take it that prosodic highlighting fulfills semantic-pragmatic or expressive functions. A group of ten prosody experts annotated the occurrences of prosodic highlighting in the corpus of study. Results confirm that prosodic highlighting is realized through an increase in pitch and duration. They differ from previous studies concerning the type of prosodic contour on occurrences of prosodic highlighting and the presence of initial secondary accent. Methodological issues are raised concerning the analysis of terminal contours and deaccenting.

MOTS-CLES : prosodie, focalisation, accentuation, contour, parole spontanée
KEYWORDS: prosody, prosodic highlighting, accent, contour, spontaneous speech

1 Introduction

1.1 Présentation de l'étude

Cette étude porte sur la réalisation phonétique et la caractérisation phonologique du phénomène de la focalisation prosodique en français. La définition de la focalisation que nous adoptons est la suivante : un soulignement prosodique de constituant (par divers moyens tels que l'accentuation ou les variations de registre et de tempo) remplissant une fonction soit sémantico-pragmatique soit emphatique. Le premier type de fonction correspond au marquage de focus (informationnel,

[1] Une partie des travaux présentés dans cette étude a déjà fait l'objet d'une publication en anglais (Godement-Berline, à paraître).

contrastif, quantificationnel/associatif, verum), et le second type correspond aux fonctions d'insistance et d'expressivité (e.g. Rossi, 1999 ; Di Cristo, 1999). Deux exemples sont donnés ci-dessous, tirés de notre corpus annoté par les experts (Fig. 1, 2).

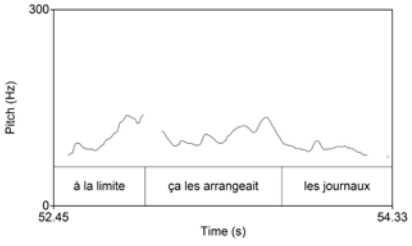


FIGURE 1 : Focalisation prosodique sur *arrangeait* (fonction : marquage de focus).

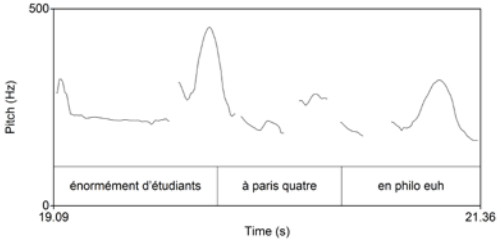


FIGURE 2 : Focalisation prosodique sur *énormément d'étudiants* (fonction : insistance).

De nombreuses études similaires ont déjà été menées (e.g. Di Cristo, 1999 ; Rossi, 1999 ; Jun et Fougeron, 2000 ; Astésano, 2001 ; Martin, 2009 ; De Looze, 2010 ; Delais-Roussarie & Di Cristo, à paraître). Cette étude se distingue par une définition fonctionnelle plus précise de la focalisation prosodique (qui est souvent limitée soit au marquage de focus soit à l'emphase) et par une méthodologie empruntée à la linguistique expérimentale faisant appel à un protocole d'évaluation explicite. Cependant, cette étude a pour inconvénient d'utiliser un corpus comportant plusieurs styles de parole différents ; ceci est dû au fait que l'expérience décrite ci-dessous s'inscrit dans un programme plus large consistant à étudier les différences de réalisation de la focalisation prosodique selon le style de parole (cf Godement-Berline, à paraître).

1.2 Etat de l'art

Un des principaux traits de la focalisation décrit par la littérature est une forte augmentation de F0, de durée et d'intensité, bien que les études précédentes ne soient pas unanimes concernant les deux derniers paramètres (e.g. Astésano, 2001 ; De Looze, 2010 ; Rossi, 1999). Un deuxième trait souvent décrit est la présence d'un accent initial sur le constituant focalisé, réalisé selon Delais-Roussarie & Di Cristo (à paraître) par une « augmentation significative de la durée de la consonne », une « glottalisation initiale si la première syllabe du mot lexical ne commence pas par une consonne » et la « présence d'une variation de hauteur perçue comme un pic mélodique extra haut ».

La présence d'un contour terminal sur le constituant focalisé est également observée, surtout concernant le marquage de focus. La définition de ce trait est toutefois problématique. Existe-t-il une forme phonologique objective de la finalité, ou bien s'agit-il d'un phénomène purement perceptif, ou qui interagit avec d'autres niveaux linguistiques (en particulier la syntaxe et la pragmatique) ? La première position est défendue notamment par Martin (2009) ou Rossi (1999), selon qui le contour terminal atteint le haut ou le bas du registre du locuteur. Cependant, certaines études (e.g. Post 2002)

montrent que des contours perçus comme terminaux peuvent ne pas correspondre à cette forme et, inversement, que des contours correspondant à cette forme peuvent ne pas être perçus comme terminaux.

Enfin, de nombreuses études sur la focalisation prosodique décrivent la présence d'une « désaccentuation » avant ou après le constituant focalisé, un trait lui aussi difficile à caractériser. Le cas le plus étudié, celui de la désaccentuation post-focale dans un énoncé d'assertion, est souvent décrit comme un contour « plat » dans le niveau bas du registre du locuteur (e.g. Rossi, 1999 ; Jun & Fougeron, 2000), éventuellement accompagné d'une accélération du débit d'articulation et d'une chute de l'intensité (De Looze 2010 ; Rossi 1999). Cette compression de la hauteur n'entraîne toutefois pas de suppression des groupements prosodiques (*dephrasing*), les frontières finales de groupe étant marquées par des allongements (e.g. Jun & Fougeron, 2000). La désaccentuation post-focale dans des énoncés à modalité interrogative avec contour terminal montant est décrite soit comme un contour montant (Rossi, 1999), soit comme un plateau de niveau haut ou moyen-haut suivi d'une montée (Jun et Fougeron, 2000), le ton final étant généralement lui aussi terminal (H%).

2 Méthodologie

2.1 Elicitation des données

Le corpus est constitué de deux extraits de parole spontanée et de huit enregistrements de locuteurs lisant à haute voix ou « interprétant » de mémoire la transcription de ces extraits spontanés (protocole RepTask ; cf Laurens et al., 2011). Le premier extrait spontané provient du corpus CID (Bertrand et al., 2008) et le second a été enregistré pour les besoins de l'expérience. Il s'agit dans les deux cas de conversations de type « bavardage » entre deux locuteurs ; dans les extraits retenus, un des deux locuteurs raconte une anecdote drôle et insolite, avec peu d'interruption de la part de l'autre. Les locuteurs sont deux hommes de la région Provence-Alpes-Côte d'Azur dans le premier extrait et deux femmes de Paris dans le second, tous âgés entre 25 et 45 ans et chercheurs ou doctorant(e)s en linguistique. Les locuteurs des versions lues et interprétées sont deux femmes et deux hommes de Paris ayant entre 20 et 40 ans et possédant tous, à des degrés divers, une expérience dans le jeu d'acteur. Leur tâche a consisté à restituer le texte comme s'ils étaient le locuteur d'origine et participaient réellement à la conversation.

Les enregistrements spontanés ont été réalisés dans une pièce calme avec un minimum de bruit de fond, à l'aide d'un micro-casque dans le premier cas et d'un micro portable Zoom H2 (format 16 bit/44.1 kHz, WAV) dans le second. Les enregistrements lus et interprétés ont été réalisés dans une chambre sourde à l'Université Paris Diderot à l'aide d'un microphone studio Rode NT1-A, d'une interface audio Roland Quad-Capture et du logiciel Audacity (format 16 bit/44.1 kHz, WAV). Le corpus (accompagné de fichiers d'alignement au format Textgrid) est disponible à l'adresse http://www.llf.cnrs.fr/reptask.

2.2 Annotation des occurrences de focalisation prosodique

Un groupe de dix experts en prosodie a annoté les occurrences de focalisation prosodique dans le corpus, en prenant pour unité d'annotation le mot. Les experts sont des chercheurs, doctorant(e)s ou

étudiant(e)s de niveau Master en phonétique, phonologie ou pragmatique ayant de l'expérience dans l'étude de la prosodie. Chaque expert a annoté quatre enregistrements, de sorte que chaque enregistrement a été annoté par quatre experts. Les experts ont écouté les enregistrements (sans limite de temps et sans l'aide d'un logiciel d'analyse acoustique permettant de visualiser le signal et ses caractéristiques prosodiques) et se sont référés pour l'annotation aux traits de la focalisation prosodiques cités plus haut (seuls ou en combinaison) : augmentation de hauteur, de durée ou d'intensité, présence d'accent initial, contour terminal sur le constituant focalisé, et désaccentuation avant ou après le constituant focalisé. Il a été demandé aux experts de ne pas annoter les accents uniquement rythmiques (marquant la frontière initiale ou finale d'un groupe prosodique). Seuls les mots annotés par au moins trois sur quatre experts ont été considérés comme des occurrences de focalisation prosodique.

2.3 Analyse prosodique

Les enregistrements ont été segmentés en mots, syllabes et phones à l'aide de l'extension de Praat EasyAlign (Goldman, 2011). Ils ont ensuite été analysés à l'aide de l'extension de Praat Prosogram (Mertens, 2004), qui donne une série de mesures pour chaque syllabe (en utilisant la segmentation préalable). Basé sur un modèle psycho-acoustique de la perception tonale, Prosogram mesure la F0 sur les portions voisées significatives uniquement, déterminées ici par le programme au sein de la rime syllabique. La F0 a été convertie en demi-tons (relatifs à 1 Hz) afin de pouvoir permettre la comparaison entre locuteurs. La durée a été normalisée par rapport à la structure syllabique en divisant la durée de chaque syllabe par le nombre de phones dans la syllabe, et par rapport au débit de parole des locuteurs en convertissant les valeurs précédentes en z-scores pour chaque locuteur.

Chaque occurrence de focalisation prosodique a été analysée visuellement et auditivement sur Praat par l'auteur afin d'en caractériser le contour prosodique, en utilisant la transcription ToBI pour le français (cf Delais-Roussarie et al., 2015). L'empan du contour au niveau du mot (i.e., sur quelle(s) syllabe(s) du constituant focalisé s'étend le contour) a également été déterminé par l'auteur. Enfin, la présence d'un accent initial sur la frontière gauche du constituant focalisé a été déterminée à l'aide de la fonction de détection automatique de proéminences du logiciel Analor (Avanzi et al., 2008), à partir de la segmentation syllabique préalable.

3 Résultats

Le corpus est composé d'un total de 5644 mots et de 7800 syllabes. Il contient 54% de mots lexicaux. Il contient 56,6% de syllabes de type CV, 14,8% de type V, 13,6% de type CVC, 9,8% de type CCV et 5,2% d'un autre type.

Le taux d'accord entre les annotations des experts a été obtenu en calculant le Kappa de Fleiss pour chaque enregistrement et en faisant la moyenne pour le corpus entier et pour chaque style de parole. Le taux d'accord pour le corpus entier est légèrement significatif (0,273, $z = 18,79$, $p = 0$). On observe des différences entre les styles de parole (taux d'accord plus haut pour la parole spontanée, suivie de l'interprétation et de la lecture).

La fréquence d'occurrence de focalisations prosodiques dans le corpus est faible (11,22 %). Elle est plus haute pour l'interprétation (14,11 %), suivie de la lecture (10,59 %) et de la parole spontanée (6,59 %) (les différences sont significatives en utilisant des intervalles de confiance à 95%).

3.1 F0 et durée syllabique moyennes

On observe une forte différence entre la F0 moyenne et la durée moyenne des syllabes comportant une focalisation et celles des autres syllabes du corpus (Fig. 3). Les données ont été analysées au moyen d'un modèle linéaire mixte, avec la présence de focalisation comme effet fixe et le locuteur, le groupe d'experts en prosodie et le style de parole comme effets aléatoires. Les p-values ont été obtenues au moyen d'un test du rapport des vraisemblances entre le modèle complet et un modèle sans l'effet fixe. Les différences sont significatives à la fois pour la F0 ($\chi^2(1) = 532{,}8$, $p < 0{,}01$) et pour la durée ($\chi^2(1) = 71{,}8$, $p < 0{,}01$).

Les différences de F0 et de durée, entre les styles de parole, des syllabes comportant une focalisation ont également été testées au moyen d'un modèle linéaire mixte avec le style de parole comme effet fixe et le locuteur et le groupe d'experts comme effet aléatoires, en utilisant à nouveau une réduction du modèle et un test du rapport des vraisemblances pour obtenir les p-values. Aucune différence ne s'avère être significative, à la fois pour la F0 ($\chi^2(2) = 0{,}137$, $p > 0{,}05$) et pour la durée ($\chi^2(2) = 3{,}617$, $p > 0{,}05$).

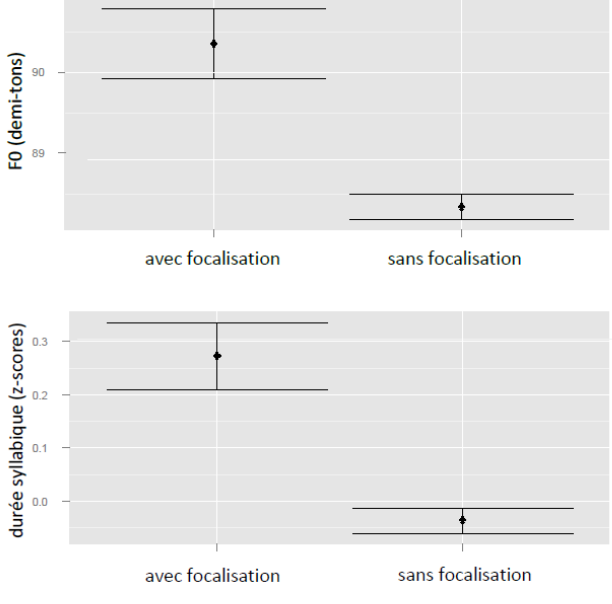


FIGURE 1: F0 et durée syllabique moyennes des syllabes comportant une focalisation prosodique et des autres syllabes du corpus

3.2 Caractérisation phonologique

On observe des différences de fréquence dans le corpus concernant le type de contour employé sur les occurrences de focalisation, le type de frontière de groupe prosodique réalisée, l'empan du contour au niveau du mot et la présence d'un accent initial. On n'observe pas de différences significatives selon ces mêmes variables entre les trois styles de parole. Ceci est confirmé par des tests du χ^2 de Pearson pour le type de contour ($\chi^2(12) = 13{,}20$, $p > 0{,}05$), le type de frontière ($\chi^2(4) = 8{,}78$, $p > 0{,}05$) et l'empan du contour ($\chi^2(6) = 7{,}78$, $p > 0{,}05$). Pour la présence d'accent initial, un modèle linéaire mixte a été employé avec le style de parole comme effet fixe et le locuteur et le groupe d'experts comme effets aléatoires, et n'a pas révélé de différence significative ($\chi^2(2) = 4{,}07$, $p > 0{,}05$).

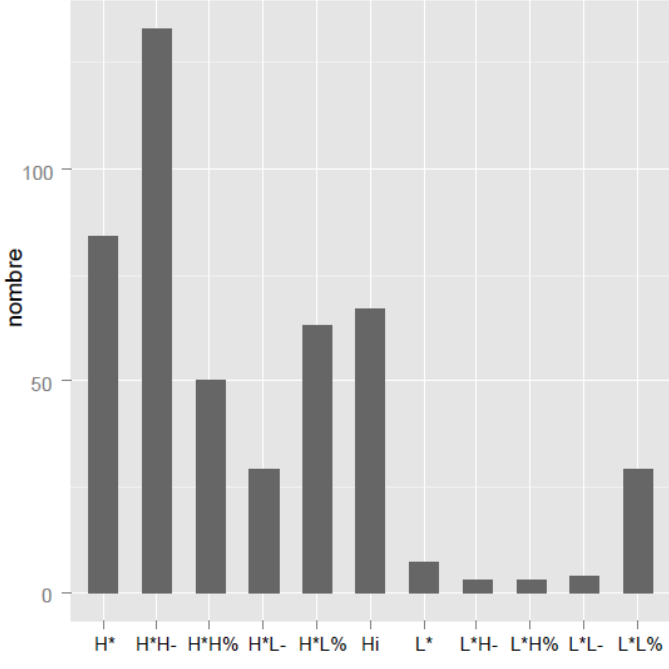


FIGURE 2 : Contours prosodiques employés sur les occurrences de focalisation

La majorité des contours sont montants (Fig. 4). Le contour le plus fréquent est H*H- (28,2 %), un contour de continuation majeure à la fin d'un groupe intermédiaire (ip), suivi de H* (17,8 %) et de Hi (14,2 %), un accent initial au début d'un groupe de mots focalisé. Les deux types de groupe prosodique les plus fréquemment réalisés à la frontière (initiale ou finale) des occurrences de focalisation sont le groupe intermédiaire (35,8 %) et le groupe accentuel (33,5 %). Lorsque le constituant focalisé fait plus d'une syllabe, les contours sont majoritairement réalisés sur la dernière syllabe (72,7 %). Une minorité de constituants focalisés plurisyllabiques comporte un accent initial (41,2 %).

4 Discussion

Les résultats confirment les études précédentes sur certains points et les contredisent sur d'autres. Les occurrences de focalisation prosodique présentent une nette augmentation de F0 et de durée syllabique, comme le notent Di Cristo (1999), Rossi (1999), Jun et Fougeron (2000), De Looze (2010) et Astésano (2001). Nous n'avons pas pu faire de mesures d'intensité car la distance micro-locuteur n'est pas fixe dans nos enregistrements (ce afin de ne pas gêner la spontanéité et l'expressivité des locuteurs). En revanche, notre caractérisation phonologique est différente de celle des mêmes auteurs, notamment car le contour prosodique le plus fréquent sur les occurrences de focalisation est le contour de continuation H*H- et car on observe une présence relativement faible d'accent initial. Le fait qu'on observe de façon minoritaire des frontières de groupe intonatif est contradictoire avec la fonction de marquage de focus (qui appelle normalement la réalisation d'un contour terminal, cf Di Cristo, 1999 ; Rossi, 1999 ; Martin, 2009) mais pas nécessairement avec les fonctions d'insistance et d'expressivité. En effet, l'insistance est décrite comme marquée principalement par la présence d'un accent initial (Di Cristo, 1999 ; Delais-Roussarie & Di Cristo, à paraître ; Astésano, 2001) ; le marquage de l'expressivité est quant à lui amalgamé avec le contour terminal (Rossi, 1999 ; Di Cristo, 1999) ou bien affecte l'énoncé entier (Di Cristo, 1999).

Plusieurs autres analyses de la focalisation pourraient être menées à partir du corpus, mais certaines présentent de sérieux obstacles méthodologiques. Comment déterminer, par exemple, la présence d'un contour terminal sur le constituant focalisé puisque, comme le montrent des études telles que Post (2002), ce contour ne possède aucune forme phonologique stable ? De même, comment objectiver la présence de désaccentuation sur la séquence précédant ou suivant le constituant focalisé ? Outre la variation de réalisation observée, il est difficile de différencier la désaccentuation de manière quantitative d'une simple baisse de hauteur (et de durée et d'intensité). Une solution possible serait de faire des mesures de différences de registre et de tempo (à l'aide par exemple des extensions de Praat ADoReVA et ADoTeVA ; cf De Looze, 2010) sur de la parole de laboratoire, par exemple sur des énoncés assertifs présentant un marquage de focus informationnel étroit, afin d'obtenir des seuils de différence à utiliser ensuite pour déterminer la présence de désaccentuation dans d'autres types de parole.

Cette étude sera prochainement complétée par une analyse des configurations tonales de la focalisation et de sa réalisation au niveau infrasyllabique (cf Astésano, 2001). Nous étudierons également l'influence de la catégorie lexicale des constituants focalisés.

5 Conclusion

Cette étude porte sur la focalisation prosodique en français, dans ses fonctions aussi bien sémantico-pragmatiques qu'emphatiques. Elle emploie une méthodologie expérimentale faisant appel à un protocole d'évaluation explicite. Nos résultats confirment que la focalisation est réalisée par une augmentation de hauteur et de durée. Ils diffèrent de la littérature précédente du point de vue de l'interprétation phonologique de la focalisation, notamment le type de contour prosodique employé et la présence d'accent initial. Une prochaine étude portera sur la réalisation prosodique des différentes fonctions de la focalisation (marquage de focus, insistance, expressivité).

Remerciements

Ce travail a bénéficié d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'Avenir » (référence : ANR-10-LABX-0083). Merci à Philippe Martin, Jean-Marie Marandin, Fabián Santiago, Barbara Hemforth, Georges Boulakia, Hiyon Yoo et Elisabeth Delais-Roussarie pour leur aide à différents stades, ainsi qu'à tous les participants.

Références

ASTESANO C. (2001). *Rythme et accentuation en français : invariance et variabilité stylistique*. Paris : L'Harmattan.

AVANZI M., LACHERET-DUJOUR A., VICTORRI B. (2008). ANALOR – A tool for semi-automatic annotation of french prosodic structure. Actes de *Speech Prosody 2008*, 119-122.

BERTRAND R., BLACHE PH., 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, 1-30.

DELAIS-ROUSSARIE E., POST B., AVANZI M., ET AL. (2015). Intonational Phonology of French: Developing a ToBI System for French. *Intonation in Romance*. Oxford: Oxford University Press, 63-100.

DELAIS-ROUSSARIE E., DI CRISTO A. (à paraître). L'accentuation. *La Grande Grammaire du français*. Arles : Actes Sud.

DE LOOZE C. (2010). *Analyse et Interprétation de l'Empan Temporel des Variations Prosodiques en Français et en Anglais*. Aix-Marseille Université.

DI CRISTO A. (1999). Le cadre accentuel du français contemporain : essai de modélisation. *Langues* 2, 258-267.

GODEMENT-BERLINE R. (à paraître). Using a replication task to study prosodic highlighting. Actes de *Speech Prosody 2016*.

GOLDMAN J.-PH. (2011). EasyAlign: an automatic phonetic alignment tool under Praat. Actes de *Interspeech 2011*, 3233-3236.

JUN S.-A., FOUGERON C. (2000). A Phonological model of French intonation. *Intonation: Analysis, Modeling and Technology*. Dordrecht : Kluwer Academic Publishers, 209-242.

LAURENS F., MARANDIN J.-M., PATIN C., YOO H. (2011). The Used and the Possible – The Use of Elicited Conversations in the study of Prosody. Actes de *IDP 2009*, 239-257.

MARTIN PH. (2009). *Intonation du français*. Paris : Armand Colin.

MERTENS P. (2004). The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model. Actes de *Speech Prosody 2004*, 2320-2323.

POST B. (2002), French tonal structures. Actes de *Speech Prosody 2002*, 583-586.

ROSSI M. (1999). *L'intonation : Le système du français*. Paris : Ophrys.

Un Corpus de Flux TV Annotés pour la Prédiction de Genres

Mohamed Bouaziz[1, 2] Mohamed Morchid[1] Richard Dufour[1]
Georges Linarès[1] Prosper Correa[2]
(1) LIA, 339 Chemin des Meinajaries, 84140 Avignon, France
(2) EDD, 28 Boulevard de Port-Royal, 75005 Paris, France
mohamed.bouaziz@alumni.univ-avignon.fr, {mohamed.morchid, richard.dufour, georges.linares}@univ-avignon.fr, pcorrea@edd.fr

RÉSUMÉ

Cet article présente une méthode de prédiction de genres d'émissions télévisées couvrant 2 jours de diffusion de 4 chaînes TV françaises structurés en émissions annotées en genres. Ce travail traite des médias de masse de flux de chaînes télévisées et rejoint l'effort global d'extraction de connaissance à partir de cette grande quantité de données produites continuellement. Le corpus employé est fourni par l'entreprise EDD, anciennement appelée "L'Européenne de Données", une entreprise spécialisée dans la gestion des flux multimédias. Les expériences détaillées dans cet article montrent qu'une approche simple fondée sur un modèle de n-grammes permet de prédire le genre d'une émission selon un historique avec une précision avoisinant les 50 %.

ABSTRACT

A Genre Annotated Corpus of French Multi-channel TV Streams for Genre Prediction.

This paper presents a method for telecast genre prediction covering 4 French TV channels on two consecutive days and structured into genre labelled telecasts. This work falls within the mass media management as regard to television channel streams and joins the global endeavours for extracting meaningful knowledge from the huge quantity of data continuously produced. The used dataset is provided by EDD (previously known as "L'Européenne de Données"), a company specialized in managing multimedia streams. Preliminary experiments conducted in this paper show that an n-gram straightforward approach allows us to predict the TV program genre according to the history with an accuracy of about 50%.

MOTS-CLÉS : Genres Télévisuels, Corpus Multi-chaîne, Structuration de Flux.

KEYWORDS: Telecast Genres, Multi-channel Corpus, Stream Structuring.

1 Introduction

La télévision a pris une place très importante dans la société avec un très grand nombre de chaînes télévisées disponibles qui n'a cessé de croître ces dernières décennies. Il apparaît alors important de pouvoir structurer ces données, les indexer, et les rendre facilement accessibles pour les professionnels ainsi que pour le grand public. Les scientifiques se sont donc fortement intéressés à l'analyse et la structuration du contenu audio dans un contexte multi-chaînes (Bredin *et al.*, 2014; Bouchekif *et al.*, 2013). Malheureusement, peu de corpus de documents audio sont disponibles, d'autant plus quand il s'agit de corpus en langue française. En outre, ces corpus, lorsqu'ils sont disponibles, sont

annotés en ne considérant qu'une seule chaîne (Amaral & Trancoso, 2003) ou en ne traitant qu'un seul aspect du problème de structuration ou une seule tâche spécifique. Il est donc difficile de les utiliser dans un contexte d'étude différent (Wang *et al.*, 2006; Gravier *et al.*, 2012). Les corpus multi-chaînes disponibles font face à deux problèmes majeurs. Premièrement, l'enregistrement est fait à des horodatages différents, ce qui rend difficile, voire impossible, la comparaison entre les différents canaux. Deuxièmement, la taille de l'ensemble des programmes pris en compte sont généralement considérablement réduites. Par exemple, récemment, le corpus REPERE (Giraudel *et al.*, 2012) contient des émissions télévisuelles provenant de différentes chaînes françaises enregistrées à des dates différentes dans la période comprise entre 2012 et 2014. Les émissions ont été extraites à partir de moments précis de la journée et ont été également choisies pour contenir principalement des genres télévisuels spécifiques localisés (journaux, débats, interviews…) excluant les autres programmes diffusés tout au long de la journée (documentaires, films, inter-programmes…).

Cet article présente, dans un premier temps, un corpus audio de flux TV de 2 jours enregistré à partir de la diffusion de 4 chaînes françaises (TF1, M6, France 5 et TV5 Monde). Ensuite, des expérimentations préliminaires d'un système temps-réel d'aide à l'annotation manuelle sont conduites en évaluant différentes représentations d'une séquence de $n-1$ émissions dans l'objectif de prédire le genre de la n^{eme} émission qui suit. Enfin, nous discutons de l'intérêt de regrouper les flux considérant leur contenu (généraliste/semi-thématique). Certains travaux se sont intéressés à la problématique de structuration de flux TV. (Poli, 2008) est le seul, à notre connaissance, à s'être appuyé sur les anciennes séquences de programmes pour structurer un flux. En apprenant un Modèle de Markov Caché Contextuel sur l'historique de diffusion de chaque chaîne, (Poli, 2008) obtient des grilles précises à 97 %. Les autres travaux utilisent strictement le contenu audio-visuel (Liang *et al.*, 2005; Wang *et al.*, 2008; El-Khoury *et al.*, 2010) ou n'exploitent l'historique d'émissions que pour des raffinements *a posteriori* (Naturel *et al.*, 2007; Manson & Berrani, 2010; Ibrahim & Gros, 2011). Le corpus présenté est généré par l'entreprise EDD, une entreprise française âgée de 35 ans et spécialisée dans la collecte, l'analyse, l'indexation et la redistribution de divers types de journaux, magazines, et données radio et TV. EDD offre également, pour des professionnels et des administrations, plusieurs services comme une plate-forme de panorama de presse, un système d'alerte thématique et un moteur de recherche multimédia. EDD a mis à la disposition du LIA des données audio enregistrées sur 2 jours de diffusion relatives à environ 4 chaînes radio et TV. Avec les données de télévision, nous avons des méta-données (chaque 2 minutes) précisant le titre de l'émission en cours de diffusion. Elles sont enrichies, pour un certain nombre de programmes, par des informations complémentaires (e. g. description, genre, etc.) et une transcription automatique produite par SPEERAL, le système de reconnaissance automatique de la parole du LIA (Linarès *et al.*, 2007).

Ce papier est structuré comme suit. La section 2 présente la taxonomie de genres. La section 3 décrit la procédure de collecte et d'annotation du corpus. La section 4 est dévolue aux expérimentations pour la prédiction du genre télévisé. Enfin, les conclusions ainsi que les travaux futurs sont présentés dans la section 5.

2 Taxonomie des Genres

La classification des émissions en genres est souvent une tâche subjective qui dépend de la perception de l'annotateur. Par conséquent, plusieurs entités compétentes, comme l'Institut National de l'Audio-visuel (INA), ont créé chacune leur propre taxonomie. L'INA dispose des archives de flux de chaînes

radio et TV françaises, dont les émissions sont classées en 52 genres différents (INA, 2002).

La conception d'une taxonomie trop exhaustive peut conduire à un certain nombre de confusions au sein de la nomenclature. En conséquence, la différence entre certains genres peut être mineure. Par exemple, les *films* sont très proches des *courts métrages* et des *téléfilms*. En outre, on ne peut que difficilement distinguer de différences entre un *talk-show* comme "On n'est pas couché" et un *magazine de débat* avec généralement un ou deux présentateurs, accompagnés d'invités (e. g. "Le magazine de la santé").

Vu le grand nombre de genres et la difficulté pour l'annotateur de les distinguer, l'identification d'un genre particulier s'avère être une tâche difficile pour un être humain et d'autant plus pour un système automatique. Nous avons alors établi une taxonomie inspirée de celle proposée par l'INA en procédant à un certain nombre de fusions permettant aux genres d'être à la fois bien définis et distincts entre eux tout en couvrant les genres d'émission les plus fréquement rencontrées. Le contenu visuel est divisé en 15 genres définis comme suit : *Inter-programmes (IP), Actualité, Météo, Dessins animés, Fiction, Documentaire, Téléachat, Plateau/Débat, Magazine de reportage, Autres magazines, Musique, Télé-réalité, Programme court, Jeu* et *Autres*.

Les *inter-programmes* sont référencés comme un genre particulier dans la mesure où ils apparaissent en général entre deux émissions ou comme pause au sein d'une même émission. Ce genre, dénommé par la suite IP, englobe les *publicités*, *jingles*, *génériques* de début et de fin de programmes, etc. Nous avons choisi de considérer les *génériques* comme des *IP* vu leurs particularités, principalement acoustiques, pouvant aider à la recherche des IP dans le cadre de la prochaine étape de ce travail, à savoir, la structuration automatique des flux TV. Quant au genre *Fiction*, il inclut les *films*, *courts-métrages*, *séries* ainsi que les *feuilletons*. Pour ce qui est de la catégorie *autres magazines*, celle-ci concerne uniquement les magazines qui ne sont ni sous forme de débat ni constitués d'une suite de reportages (e. g. Dr CAC, Astuces du Chef...). Les programmes moins fréquents et les émissions couvrant un événement particulier comme les *événements sportifs* et *les émissions de variétés* sont rassemblés dans la catégorie *autres*.

FIGURE 1: Un exemple d'une courte séquence de programmes sur TF1.

3 Le Corpus Multi-chaînes

L'utilisation du contenu, dans le cadre des travaux de structuration de flux TV cités dans l'introduction, se focalise essentiellement sur l'information vidéo. Ce corpus se limite à l'information audio vu que nous comptons, dans la suite de ce travail, exploiter l'audio pour la structuration des flux TV. L'utilisation de ce média, nous permettra ensuite d'étendre notre futur système aux flux radio également traités par EDD. Cette section décrit le processus de collecte ainsi que le processus d'annotation de ces données. Le contexte et les statistiques du corpus sont ensuite détaillés. Une discussion sur les catégories de programmes est finalement délivrée.

3.1 Annotation

L'objectif principal de la constitution de ce corpus annoté manuellement en genres, est de concevoir un environnement digital homogène permettant d'expérimenter les représentations séquentielles de programmes télévisés et assurant une mission d'aide à l'annotation manuelle. Le corpus TV est segmenté, non seulement en émissions et inter-programmes, mais également en considérant comme différents segments les parties de programmes séparées par des pauses représentées par le genre *inter-programmes*.

La figure 1 présente un exemple d'une séquence de programmes annotés. La première étape du processus d'annotation manuelle consiste à tirer profit des méta-données relatives à chaque tranche de 2 minutes de retransmission afin d'établir une structure primaire concernant les frontières entre les différentes émissions. En effet, ces méta-données ne sont ni précises ni exhaustives. De plus, elles n'offrent aucune information en ce qui concerne le *timing* des *inter-programmes*. Ces circonstances ont rendu l'annotation manuelle très difficile sachant que nous traitons des données audio.

Concrètement, l'outil libre *WaveSurfer* (Sjölander & Beskow, 2000) de visualisation et de manipulation de son a été utilisé. En procédant à une annotation manuelle effectuée par un seul annotateur, un ensemble d'informations est affecté à chaque segment, *i.e.* l'instant de début et de fin, le genre correspondant provenant de notre taxonomie et, si possible, le nom de l'émission correspondante audit segment. Ces labels sont enregistrés dans un fichier texte pour chaque tranche de 24 heures de flux.

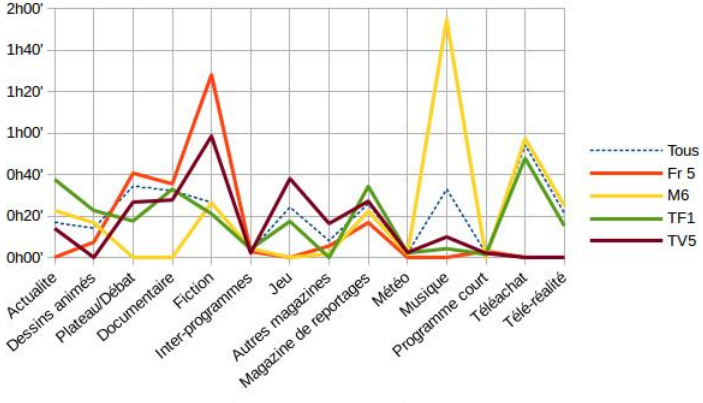


FIGURE 2: La durée moyenne des segments par genre

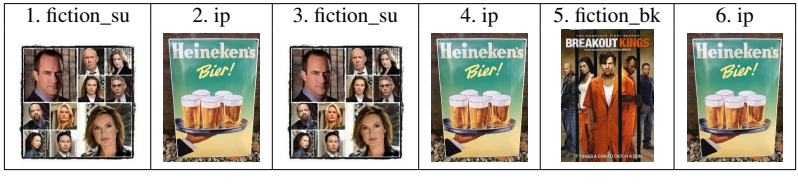


FIGURE 3: Un exemple d'ambiguïté pour deux fictions successives : Special Unit (su) et Breakout Kings (bk) sur TF1. *ip : inter-programmes*

Un problème qui se pose est que certaines émissions peuvent combiner deux genres considérés assez éloignés l'un de l'autre. Par exemple, le programme appelé "Petits secrets entre voisins" de la chaîne TF1 associe à la fois le genre *fiction* ainsi que le genre *télé-réalité* [1]. Dans de telles situations, la liste des genres concernés est spécifiée en les concaténant et en les séparant par le signe **+**.

Par ailleurs, les génériques des émissions sont considérés comme des *inter-programmes*. Ceci dit, une ambiguïté se pose toujours pour discerner certains types de génériques (cf. première colonne du tableau 1). La question de les joindre au segment d'*inter-programmes* ou à celui de l'émission voisine se pose. Il a été décidé de créer un label différent pour chaque cas comme montré dans le tableau 1.

Cas	Étiquette
Parole non répétée, sur la musique du générique	générique_parole_non_répétée
Parole répétée, sur la musique du générique	générique_parole_répétée
Parole introductive suivie du générique de début & générique de fin suivi d'une parole de fin	parole_extra_générique

TABLE 1: Annotation des génériques ambigus

3.2 Détails et Statistiques

Le corpus fourni par l'entreprise EDD est composé de retransmissions de 4 chaînes TV françaises (TF1, M6, France 5 et TV5 Monde) pour une période de 2 jours entre le 10 et 12 février 2014.

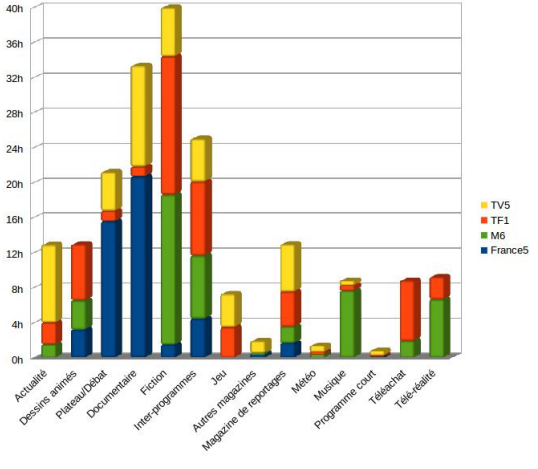


FIGURE 4: *Distribution des genres dans le corpus.*

Comme illustré dans la figure 4, les 4 chaînes TV ne partagent pas la même distribution de genres, chose prévisible sachant que ces chaînes grands publics ne partagent pas une même politique éditoriale. Pour les deux chaînes généralistes (TF1 et M6), la *fiction* est le genre de loin le plus représenté avec quasiment un tiers de parts de diffusion. En outre, ces deux chaînes sont les seules qui retransmettent des émissions de *télé-réalité* et, étant donné qu'elles sont privées, du *téléachat*. Pour cette raison,

1. Ce type de programme TV est désigné récemment comme "réalité scénarisée" ou *scripted reality*

Genres	Fr5	TV5	TF1	M6	Gen.	Total
Actualité	0	38	4	4	4	46
Cartoon	26	0	17	12	14.5	55
Plateau/Débat	23	10	4	0	2	37
Documentaire	35	25	2	0	1	62
Fiction	1	6	45	39	42	91
Inter-prog.	95	134	124	101	112.5	454
Jeu	0	6	12	0	6	18
Autres mag.	4	5	0	5	2.5	14
Mag. de rep.	6	12	7	5	6	30
Météo	0	18	8	15	11.5	41
Musique	0	3	9	4	6.5	16
Prog. court	3	15	6	0	3	24
Téléachat	0	0	1	2	1.5	3
Télé-réalité	0	0	10	16	13	26
Total	193	272	249	203	226	917

TABLE 2: Nombre de segments pour chaque genre. *FR5 : France 5, Gen. : Moyenne des chaînes généralistes, Inter-prog. : Inter-programmes, Mag. : Magazine, Rep. : Reportages, Prog. : Programme.*

elles disposent des plus grandes parts d'*inter-programmes* parmi les 4 chaînes. En effet, nous pouvons remarquer à partir de la figure 2 que les segments de *fiction* durent environ deux fois plus longtemps dans les chaînes plus spécialisées France 5 et TV5, deux chaînes semi-thématiques, que dans TF1 et M6. De plus, TV5 possède le plus grand nombre d'*inter-programmes*, comme le montre le tableau 2. Cependant, cela ne signifie pas forcément qu'elle possède la plus grande quantité d'*inter-programmes* vu le nombre relativement important d'émissions de courtes durées, comme les *journaux d'actualité*, les *prévisions météo* et *programmes courts*.

Malgré la similarité significative entre les politiques des deux chaînes généralistes, nous pouvons tout de même constater quelques différences. Par exemple, M6 consacre, tard dans la nuit, une portion de retransmission d'environ 4 heures par jour à la *musique*. D'un autre côté, TF1 se distingue par diverses émissions de jeu. Pour ce qui est des deux chaînes semi-thématiques, France 5 est spécialisée dans les *documentaires* et alloue aussi une part représentant quasiment un tiers de la programmation télévisuelle aux émissions de *débat*. Néanmoins, elle ne diffuse ni des *journaux* ni des *prévisions météo*. Ces deux derniers genres constituent par contre le thème principal de TV5 Monde.

4 Expériences

L'objectif de cet article est de construire un système d'aide à l'annotation manuelle. Ce dernier est évalué lors d'expériences de classification automatique en genres des programmes TV, présentées dans cette section, qui sont conduites sur le corpus TV multi-chaînes décrit dans la section 3. Le but de cette tâche est de prédire le genre d'une émission en cours, sachant les émissions qui la précèdent, à l'aide d'un modèle de n-grammes proposé dans la section 4.1, et d'évaluer ensuite la consistance de l'utilisation de plusieurs chaînes durant le processus de prédiction de genres.

4.1 La Prédiction du Genre

Le but de cette expérience est de découvrir le genre d'un programme télévisé en cours sachant les programmes précédents. Cette étude emploie un modèle n-gramme (Brown *et al.*, 1992) évaluant la probabilité *a posteriori* du $i^{\text{ème}}$ programme x_i étant donné les $n-1$ programmes précédents. Ceci équivaut à évaluer la probabilité d'observer un genre particulier dans le contexte d'un historique restreint aux genres des $n-1$ émissions précédentes.

$$p(x_i|x_{i-(n-1)},\ldots,x_{i-1}) = \frac{count(x_{i-(n-1)},\ldots,x_{i-1},x_i)}{count(x_{i-(n-1)},\ldots,x_{i-1})} \quad (1)$$

La figure 3 illustre un exemple d'une séquence de deux émissions successives. Il est admis que l'étiquetage automatique des programmes n'est pas une tâche triviale. D'un côté, un modèle n-gramme (avec $n > 3$) va considérer les segments de deux *fictions* successives comme appartenant à une seule émission. D'un autre côté, si nous prenons en compte le label "inter-programmes" dans le calcul des probabilités (*i.e.* comme un genre d'émission de même niveau que les autres) et vu le nombre relativement grand des segments de *fiction*, un modèle n-gramme appris sur un historique assez court ($n < 3$) va produire uniquement une suite de *fictions* et d'*inter-programmes* (cf. extrait de programmation télévisée dans la figure 3).

4.2 Protocole Expérimental

Afin de simplifier cette tâche, des choix ont été effectués. D'une part, nous avons pris en compte uniquement les premiers genres de chaque label. D'autre part, nous avons joint les occurrences des trois sortes de génériques ambigus (cf. tableau 1) au segment d'*inter-programmes* voisin. Nous avons donc spécifié un ensemble de configurations ayant les même données de test, à savoir, la deuxième journée de TF1, et dépendant des paramètres suivants (cf. tableau 3) :

- La taille du modèle n-gramme (de 3-grammes à 6-grammes),
- Les données utilisées lors de l'apprentissage du modèle de n-gramme,
- La prise en compte des *inter-programmes* au sein des données d'apprentissage et de test.

La première phase consiste à apprendre un modèle n-gramme en se restreignant à la première journée de TF1. Ensuite, une première expérience de la modélisation multi-chaînes est réalisée en combinant les données d'apprentissage de plusieurs chaînes. En ajoutant les deux journées de M6 à la première journée de TF1, nous voulons vérifier, sachant qu'il s'agit de deux chaînes concurrentes possédant des lignes éditoriales proches, si un modèle combiné pourra être plus performant. Enfin nous essayons de pousser encore le caractère multi-flux en combinant les deux chaînes restantes, à savoir France 5 et TV5 Monde, malgré leur divergence présumée par rapport aux deux autres chaînes généralistes.

4.3 Résultats

En entraînant le modèle sur la première journée de TF1, la meilleure performance est obtenue en employant les modèles 5-grammes et 6-grammes, et en excluant les *inter-programmes*. Jusqu'alors, le fait de ne pas inclure les *inter-programmes* était bénéfique, ceci étant probablement dû au fait que l'on enlève impact négatif qu'ils peuvent produire sur la taille effective de l'historique (les

inter-programmes sont présents entre chaque deux segments successifs de programmes) et de la masse de probabilité qu'ils peuvent monopoliser (particulièrement en ce qui concerne les probabilités uni-grammes). Ensuite, nous observons la manière dont un modèle multi-chaîne pourrait influer le comportement du système. Malheureusement, au lieu de renforcer le modèle n-gramme, il semble que l'ajout des autres chaînes, mais aussi l'exclusion des *inter-programmes*, y insère de la "pollution". Ceci est probablement dû à une divergence importante entre les chaînes utilisées. Dans la suite de ce travail, nous comptons étudier plus en détail ces résultats et conduire davantage d'expériences afin d'acquérir une meilleure compréhension de la tâche de prédiction de genres.

Taille	Inter-prog	Appr.	Précision	Rappel	F-mesure	TR
3g & 4g	inclus	TF1 1/2	0.473	0.484	0.478	0.169
5g			0.481	0.433	0.456	0.381
6g			0.489	0.551	0.518	0.195
3g	non inclus		0.537	0.462	0.497	0.367
4g			0.579	0.508	0.541	0.350
5g & 6g			**0.588**	**0.513**	**0.548**	**0.333**
3g	inclus	TF1 1/2 + M6	0.406	0.401	0.403	0.305
4g			0.503	0.447	0.473	0.280
5g			0.441	0.453	0.447	0.331
6g			0.476	0.430	0.452	0.280
3g	non inclus		0.402	0.363	0.382	0.383
4g			0.408	0.305	0.349	0.467
5g			0.408	0.305	0.349	0.467
6g			0.410	0.322	0.360	0.450
3g	inclus	TF1 1/2 + M6 + Fr5 + TV5	0.343	0.356	0.349	0.220
4g			0.379	0.384	0.382	0.297
5g			0.321	0.370	0.344	0.305
6g			0.330	0.3910	2.000	0.288
3g	non inclus		0.317	0.343	0.329	0.450
4g			0.311	0.256	0.281	0.667
5g			0.326	0.288	0.306	0.550
6g			0.335	0.238	0.279	0.650

TABLE 3: Précisions, Rappels, F-mesures et Taux de Réussite (TR) observés pour différentes configurations d'apprentissage. *1/2* : jour 1, *inter-prog.* : inter-programmes, *Appr.* : données d'apprentissage.

5 Conclusion

Nous avons présenté dans cet article un corpus télévisé s'étalant sur une période de 2 jours ainsi que des méthodes fondées sur les n-grammes permettant de prédire l'étiquette d'une émission sachant les programmes précédents. Ce flux multi-chaîne a été, dans un premier temps, annoté manuellement avec le genre des émissions en utilisant une taxonomie que nous avons conçue auparavant. Diverses statistiques et caractéristiques de ce corpus ont été décrites dans ce travail. Par ailleurs, nous avons proposé un système préliminaire temps-réel d'aide à l'annotation manuelle de prédiction de genres d'émissions. Le meilleur modèle n-gramme a obtenu des résultats prometteurs en ce qui concerne la tâche de prédiction de genres malgré quelques déceptions. Les travaux futurs se concentreront sur les détails de l'expérience de prédiction de genre et établiront, dans une prochaine étape, un système temps-réel de structuration automatique des flux TV.

Références

AMARAL R. & TRANCOSO I. (2003). Topic indexing of tv broadcast news programs. In *Computational Processing of the Portuguese Language*, p. 219–226. Springer.

BOUCHEKIF A., DAMNATI G. & CHARLET D. (2013). Complementarity of lexical cohesion and speaker role information for story segmentation of french tv broadcast news. In *Statistical Language and Speech Processing*, p. 51–61. Springer.

BREDIN H., LAURENT A., SARKAR A., LE V.-B., ROSSET S. & BARRAS C. (2014). Person instance graphs for named speaker identification in tv broadcast. In *Proceedings of Odyssey*.

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.

EL-KHOURY E., SÉNAC C. & JOLY P. (2010). Unsupervised segmentation methods of tv contents. *International Journal of Digital Multimedia Broadcasting*, **2010**.

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.

IBRAHIM Z. A. A. & GROS P. (2011). Tv stream structuring. *ISRN Signal Processing*, **2011**.

INA (2002). Connaissance des fonds, histoire des programmes : par genre. **2**.

LIANG L., LU H., XUE X. & TAN Y.-P. (2005). Program segmentation for tv videos. In *Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on*, p. 1549–1552 : IEEE.

LINARÈS G., NOCÉRA P., MASSONIE D. & MATROUF D. (2007). The lia speech recognition system : from 10xrt to 1xrt. In *Proceedings of the 10th international conference on Text, speech and dialogue*, p. 302–308 : Springer-Verlag.

MANSON G. & BERRANI S.-A. (2010). Automatic tv broadcast structuring. *International journal of digital multimedia broadcasting*, **2010**.

NATUREL X., GRAVIER G. & GROS P. (2007). Fast structuring of large television streams using program guides. In *Adaptive Multimedia Retrieval : User, Context, and Feedback*, p. 222–231. Springer.

POLI J.-P. (2008). An automatic television stream structuring system for television archives holders. *Multimedia systems*, **14**(5), 255–275.

SJÖLANDER K. & BESKOW J. (2000). Wavesurfer-an open source speech tool. In *INTERSPEECH*, p. 464–467.

WANG J., DUAN L., LI Z., LIU J., LU H. & JIN J. S. (2006). A robust method for tv logo tracking in video streams. In *Multimedia and Expo, 2006 IEEE International Conference on*, p. 1041–1044 : IEEE.

WANG J., DUAN L., LIU Q., LU H. & JIN J. S. (2008). A multimodal scheme for program segmentation and representation in broadcast video streams. *Multimedia, IEEE Transactions on*, **10**(3), 393–408.

Disfluences dans le vieillissement "normal" et la maladie d'Alzheimer : indices segmentaux, suprasegmentaux et gestuels

Diane Caussade[1, 2] Nathalie Vallée[1], Nathalie Henrich Bernardoni[1], Jean-Marc Colletta[2], Silvain Gerber[1], Frédérique Letué[3] & Marie-José Martinez[3]
(1) Univ. Grenoble Alpes, GIPSA-lab, F-38000 Grenoble, France
CNRS, GIPSA-lab, F-38000 Grenoble, France
(2) Univ. Grenoble Alpes, LIDILEM, F-38000 Grenoble, France
(3) Univ. Grenoble Alpes, LJK, F-38000 Grenoble, France
CNRS, LJK, F-38000 Grenoble, France

diane.caussade@gipsa-lab.fr, nathalie.vallee@gipsa-lab.fr, nathalie.henrich@gipsa-lab.fr, jean-marc.colletta@u-grenoble3.fr, silvain.gerber@gipsa-lab.fr, frederique.letue@imag.fr, marie-jose.martinez@iut2.upmf-grenoble.fr

RESUME

L'objectif de cette étude est d'analyser et comparer les productions langagières dans leur multimodalité de 10 personnes atteintes de la maladie d'Alzheimer (MA) appariées à 10 contrôles. Différentes mesures aux niveaux segmental et suprasegmental – erreurs, pauses et allongements vocaliques – ont été réalisées dans une tâche de répétition avec ou sans gestes imposés pour caractériser une disfluence, typique de la MA, puis observées en lien avec les gestes manuels produits. Les résultats montrent la diminution significative de la fluence chez les personnes atteintes de la MA, avec davantage d'erreurs produites au niveau lexical par le groupe Patient et au niveau phonétique par les patients au stade modéré de la maladie, ainsi que de nombreuses pauses silencieuses précédant ou suivant souvent les erreurs produites au niveau segmental. De plus, dans la tâche avec gestes imposés, la répétition de ceux-ci a impacté la fluence des groupes Contrôle et Patient avec une augmentation significative des disfluences au niveau suprasegmental et des erreurs phonétiques au niveau segmental.

ABSTRACT

Disfluencies in "normal" aging and Alzheimer's disease: segmental, suprasegmental and gestural markers

The aim of the study is to analyze and compare multimodal language productions by 10 persons with Alzheimer's disease (AD) matched to 10 controls. First, different measures at the segmental and suprasegmental levels – errors, pauses and vocalic lengthenings – have been conducted in a repetition task with or without imposed gestures to characterize a disfluency, typical of AD, and then observed in link with the production of manual gestures. Results show a significant diminution of the AD participants' fluency: more lexical errors were obtained by the patient group and at the phonetic level among patients with moderate cognitive impairment, as well as numerous silent pauses often preceding or following the production of errors at the segmental level. Moreover, a significant augmentation of suprasegmental disfluencies and phonetic errors is observed in the task with imposed gestures impacting controls and patients' fluency.

MOTS-CLÉS : Alzheimer, disfluences, segmental, suprasegmental, gestualité, aphasie, apraxie
KEYWORDS: Alzheimer, disfluencies, segmental, suprasegmental, gestuality, aphasia, apraxia

1 Introduction

En raison de l'allongement de l'espérance de vie et de la prévalence exponentielle des troubles neurocognitifs (TNC) après 65 ans, la prise en charge de ces troubles représente un défi de santé publique (Amieva et al., 2014). La théorie actuelle qui prévaut en neurosciences sépare les maladies neurodégénératives du vieillissement "normal", tout en considérant statistiquement que la majorité d'une classe d'âge avancée peut souffrir d'une maladie neurodégénérative (Albert, 2011). L'étiologie la plus fréquente des pathologies responsables de TNC est la maladie d'Alzheimer (MA). Sa cause est toujours inconnue, son traitement est actuellement symptomatique et pas encore curatif (Amieva et al., 2014). Cette maladie neurodégénérative est caractérisée cliniquement par des troubles des fonctions exécutives, mnésiques, spatio-temporelles, gnosiques, praxiques et phasiques pouvant notamment impacter les capacités communicatives (*ibid*.). Les troubles du langage font partie des premiers signes cliniques de la maladie (*ibid*). Il est à noter que les troubles dus à la MA peuvent également être associés à d'autres troubles dus au vieillissement dit "normal".

Avec le vieillissement, les capacités auditives se dégradent, ainsi que les capacités de compréhension et d'identification de la parole (Füllgrabe & Moore, 2014). La production de la parole est également altérée au niveau acoustique tel que la f0, ainsi qu'au niveau articulatoire (Schötz, 2006). Des études ont également montré une détérioration des fonctions cognitives dans le vieillissement "normal" (Amieva et al., 2014). Ainsi, la définition du vieillissement dépend d'un nombre important de facteurs intrinsèques (physiologiques et psychologiques) et extrinsèques (contextuels et environnementaux), et est généralement catégorisée, en recherche, en fonction de l'âge chronologique (Lee, 2012). Au vue de ces éléments, déterminer une appréciation normative du vieillissement est difficile d'autant plus qu'il s'accompagne de divers troubles pouvant être confondus avec les symptômes dus à des maladies telle que la MA, notamment lors du stade léger de la maladie (*ibid.*).

Caractériser les capacités de communication des personnes atteintes de la MA pourrait servir au diagnostic de cette maladie – notamment en définissant plus précisément les différences entre vieillissement "normal" et pathologique –, ainsi qu'au développement d'une prise en charge non-médicamenteuse telle qu'orthophonique.

1.1 Indices segmentaux et suprasegmentaux dans le vieillissement "normal" et la MA

Dans la MA, l'aphasie – perturbation du code linguistique affectant la production et/ou la compréhension – est marquée dès le stade léger de la maladie au niveau lexico-sémantique, notamment par les phénomènes d'anomie ou manque du mot et de paraphasie – qui consiste à la production d'un mot ou d'un phonème pour un autre (Amieva et al., 2014 ; Lee, 2012). L'apraxie de la parole – trouble neurologique, et non moteur ou sensitif, de la programmation des mouvements articulatoires de la parole – entraîne des troubles au niveau phonético-phonologique. La praxie est généralement considérée comme bien préservée jusqu'au stade sévère de la maladie (Aubin & Le Gall, 2003), bien qu'un nombre grandissant d'études tendent à remettre en question ce point de vue (Luchesi Cera et al., 2013 ; Gayraud et al., 2011). Aphasie et apraxie sont souvent des troubles associés dans la MA (Luchesi Cera et al., 2013). Avec l'intensification des troubles, des travaux récents sur la parole spontanée ont décrit de plus nombreuses disfluences – mesurables aux niveaux des pauses, des allongements vocaliques, des répétitions, … – dans la production langagière des personnes atteintes de la MA que des personnes "saines" (Lee, 2012).

L'idée reçue a longtemps été que les capacités linguistiques font partie des fonctions cognitives les plus résistantes au vieillissement "normal" (Albert, 2011). Les aspects lexicaux ne déclineraient guère et pourraient même s'améliorer avec l'âge (Joubert & Le Rouzo, 2000). Cependant les aspects phonético-phonologiques du langage n'ont été que très peu étudiés. Certaines études ont montré que la fluence

resterait intacte chez les personnes âgées (ex. : Penny et al., 1996), alors que d'autres ont mis en exergue une disfluence (ex. : Schötz, 2006). La majorité des études attribuent l'augmentation des disfluences – c'est-à-dire l'ensemble des phénomènes temporaux et verbaux comme les pauses, les hésitations, les autocorrections, … (Lee, 2012) – à un trouble de la récupération dans le vieillissement "normal" et dans la MA, soit de nature lexico-sémantique (Tran et al., 2011), soit de nature phonologique (Beeson et al., 1997). D'autres (ex. : Joubert & Le Rouzo, 2000) expliquent ce phénomène dans le vieillissement "normal" par une 'élaboration linguistique' plus fine du discours oral chez les personnes âgées que jeunes.

1.2 Indices gestuels dans le vieillissement "normal" et la MA

Même si de nombreuses études s'intéressent aux apraxies bracchio-manuelles (cf. Aubin & Le Gall, 2003 pour une revue), peu d'études s'intéressent aux gestes manuels dans le vieillissement "normal" et la MA. Pourtant le geste est étroitement lié au langage aux niveaux réceptif – aussi bien par rapport à la compréhension (Hubbard et al., 2002) qu'à la mémorisation du message verbal (Tellier, 2009) – et productif (Feyereisen et al., 2007 ; McNeill, 1992). De ce fait, observer le geste conjointement à la parole peut renseigner sur les capacités communicatives et cognitives des locuteurs, ainsi que contribuer au débat de production du langage. En effet, plusieurs hypothèses de génération sont débattues. Pour McNeill (1992) gestes manuels et parole seraient deux aspects d'un même système. Alors que pour d'autres auteurs comme Feyereisen et collaborateurs (2007), ils dépendraient de différents sous-systèmes selon leur fonction.

Des études ont montré qu'en perception, les gestes manuels aideraient les personnes atteintes de la MA à la compréhension du message verbal (ex. : Hubbard et al., 2002). Certaines études montrent qu'aux stades légers et modérés de la maladie, le taux de gestes manuels produits ne serait pas différent de celui des contrôles (Schiaratura et al., 2015 ; Hubbard et al., 2002) : les personnes atteintes de la MA produiraient significativement plus de gestes représentationnels que de gestes non-représentationnels par rapport aux contrôles. Concernant les études s'intéressant à l'effet de l'âge sur la production spontanée de gestes manuels, plusieurs études (Feyereisen et al., 2007 ; Morsella & Krauss, 2004) ont trouvé un effet significatif de l'âge au niveau du taux de gestes représentationnels, mais non au niveau des gestes non-représentationnels.

En ce qui concerne les disfluences produites en langue maternelle, McNeill (1992) a souligné l'absence de gestes produits lors de disfluences. Pourtant, selon l'hypothèse LRF (*Lexical Retrieval Facilitation*), la production de gestes manuels faciliterait l'accès phonologique de la forme des mots contenus dans le lexique mental et aiderait à la compréhension de l'intention communicative du locuteur (De Ruiter, 2009 pour une revue). Des arguments en faveur de cette hypothèse sont 1/ que le fait d'empêcher la production de gestes manuels interfère sur l'accès phonologique de la forme du mot en créant davantage de phénomènes de MBL (Mot sur le Bout de la Langue) et ralentit le débit de la parole, 2/ que la production de gestes manuels faciliterait l'accès lexical (Tellier, 2009).

La disfluence a été étudiée grâce à des tests orthophoniques de dénomination tels que la Batterie d'Evaluation des Troubles Lexicaux (ex. : Tran et al., 2011) ou de parole spontanée (ex. : Lee, 2012). Dans ce contexte, étudier la disfluence dans une tâche de répétition devrait permettre d'éviter des effets d''élaboration linguistique' et de pouvoir ainsi attribuer les disfluences à un trouble de la récupération. A notre connaissance aucune étude n'a traité les phénomènes d'hésitation dans une perspective multimodale dans les productions langagières des personnes atteintes de la MA et des personnes âgées "saines". Pourtant la prise en compte des niveaux segmental et suprasegmental conjointement aux gestes manuels pourrait apporter des éléments quant aux capacités communicatives et cognitives des locuteurs.

Le but de cette étude est d'examiner les troubles de la fluence chez les personnes atteintes de la MA. En nous basant sur les résultats d'une précédente étude pilote (Caussade et al., 2015), nous faisons l'hypothèse d'une diminution plus importante de la fluence due à la MA par rapport au

vieillissement dit "normal" pouvant s'observer aux niveaux segmental et suprasegmental avec une augmentation des erreurs et des pauses, conjointement à un taux de gestes manuels spontanément produits plus important pour un groupe Patient comparé à un groupe d'individus Contrôle, et à un taux de gestes manuels répétés moins important pour le groupe des patients. Nous faisons également l'hypothèse d'un impact de la répétition de gestes manuels sur la fluence au niveau verbal pouvant s'observer par une augmentation de la fluence aux niveaux segmental et suprasegmental dans la tâche avec gestes imposés pour les participants atteints de la MA. Pour ce faire, nous avons développé un protocole original présenté ci-dessous.

2 Matériel et méthode

2.1 Participants

Les données ont été recueillies auprès d'un groupe Patient de 10 femmes diagnostiquées par nos partenaires hospitaliers comme atteintes de TNC probablement dus à la MA et d'un groupe Contrôle apparié en sexe et en âge. Toutes les participantes sont de langue maternelle française et droitières. Selon le MMSE (Folstein et al., 1975), qui permet une évaluation clinique des TNC, notre groupe Patient présente une atteinte neurocognitive légère (19/30 < MMSE < 24/30) à modérée (10/30 < MMSE < 18/30), alors que les contrôles ne présentent pas de TNC (29/30 < MMSE < 30/30). Le t-test a permis de vérifier que les patientes sont bien appariées en âge aux contrôles et que la différence de MMSE entre les deux groupes est significative (cf. Table 1).

Nombres d'études (par ex. : Lee, 2012 ; Stern et al., 1994) ayant montré que le niveau socioéducatif (NSE) a une influence sur les fonctions cognitives dans le vieillissement pathologique et non-pathologique, la grille de Poitrenaud (Kalafat et al., 2003) a été utilisée pour évaluer le NSE des participantes selon quatre niveaux : pas de diplôme (Niveau 1), ou scolarité allant au maximum jusqu'à la fin de la classe de 4ème (Niveau 2), scolarité allant au maximum jusqu'à la fin de la classe de terminale sans obtention du baccalauréat (Niveau 3), réussite à un examen de niveau baccalauréat ou plus (Niveau 4). Le test de Student a mis en exergue le fait que le groupe Contrôle n'était pas apparié en NSE au groupe Patient (p = 0,001) (cf. Table 1), le recrutement difficile des participants n'ayant pas permis leur sélection. Par conséquent nous prêterons d'autant plus attention aux effets de cette variable explicative dans l'analyse de nos données.

	Patientes, n = 10			**Contrôles, n = 10**			
	Moyenne	**Ecart-type**	**Etendue**	**Moyenne**	**Ecart-type**	**Etendue**	**p**
Age	81,6	9,82	67-91	80,8	10,40	63-90	n.s.
MMSE	18,8	4,92	11-24	29,3	0,48	29-30	< 0,0001
NNSE	2,6	1,07	1-4	3,9	0,32	3-4	= 0,001

TABLE 1 : Caractéristiques des participantes.

2.2 Corpus et protocole

Les données audio-visuelles ont été recueillies lors de deux tâches de répétition dont une avec gestes manuels imposés et l'autre sans gestes manuels imposés. Pour ce faire, nous nous sommes basés sur les travaux de Miller (1956) sur l'empan mnésique – c'est-à-dire le nombre d'items (dans l'expérience de Miller, il s'agit d'une liste de chiffres monosyllabiques) qu'un sujet peut mémoriser et qui serait en moyenne pour un adulte de 7 ±2 – et sur la théorie du *Dual Coding* ou 'Calepin visuo-spatial' de Baddeley (1992) – qui stipule qu'encoder une information de manière multimodale par la parole et le geste renforcerait l'apprentissage. Ainsi, chacune des deux tâches a été constituée de deux comptines de 6 vers de 8 syllabes chacun composé au total de 147 lexèmes à répéter et de 453 phonèmes à articuler par participant. En ce qui concerne les gestes imposés, ils consistaient en 6

gestes représentationnels dont 4 iconiques et 2 déictiques par comptine, soit un geste par vers : ce qui fait un total de 12 gestes à répéter par participant.

Dans un premier temps, toutes les informations nécessaires à l'appariement du groupe expérimental et du groupe Contrôle ont été renseignées (i.e. NSE, MMSE). Dans un second temps, l'expérimentatrice demandait aux participants pour la tâche sans gestes imposés de « Répéter ce que je dis. » et pour la tâche avec gestes imposés de « Répéter ce que je dis et ce que je fais avec mes mains. » pour chacun des vers. Afin d'éviter toute convergence phonétique de la part de l'expérimentatrice, celle-ci répétait les productions audio-visuelles enregistrées préalablement et diffusées grâce à un ordinateur portable hp EliteBook. Les enregistrements audiovisuels ont été effectués à l'aide d'un caméscope sur pied Sony Handycam HDR-XR500 et d'un micro-cravate Audio Technica ATR35S. Afin que la situation d'expérimentation soit la plus écologique possible, les enregistrements ont été effectués dans une pièce familière aux participants.

2.3. Matériel et analyses

Ces enregistrements ont été transcrits manuellement aux niveaux segmental et suprasegmental en utilisant le logiciel Praat®, et au niveau des gestes manuels co-verbaux avec le logiciel ELAN®.

2.3.1. Aux niveaux segmental et suprasegmental

Les disfluences codées ont été relevées au niveau segmental selon le nombre d'erreurs produites et leur type. Pour ce faire, nous nous sommes basés sur la typologie utilisée par Luchesi Cera et collaborateurs (2013), et complétée :

- Ajout : insertion d'un mot, d'une syllabe ou d'un phonème ;
- Inversion : inversion entre deux mots, deux syllabes ou deux phonèmes ;
- Omission : omission d'un mot, d'une syllabe ou d'un phonème ;
- Substitution : remplacement d'un mot, d'une syllabe ou d'un phonème ;
- Troncation : troncation d'énoncé ;
- Autocorrection : correction spontanée d'erreurs articulatoires ;
- Essai-erreur : recherche du point articulatoire d'un phonème ou d'une séquence de phonèmes dans la tentative de produire le mouvement correct ;
- Répétition : production plus d'une fois d'un mot, d'une syllabe ou d'un phonème.

Nous situant du côté des études qui font l'hypothèse que les disfluences sont attribuées à un trouble de la récupération, nous nous sommes également intéressés aux types d'erreurs selon le niveau lexical ou phonologique.

Les disfluences codées au niveau suprasegmental ont été relevées à partir du nombre de pauses et d'allongements vocaliques produits. Nous avons utilisé dans cette étude la typologie employée par Lee (2012) :

- Pauses silencieuses : trois types de phases silencieuses apparaissant au sein d'une frontière syntaxiques ont été annotées : < 200 ms, [200 ms,1000 ms] et > 1000 ms ;
- Pause sonore : phrase constituée d'un item quasi-lexical ('euh' en français) qui entrecoupe une succession de phonèmes ;
- Allongement vocalique : allongement de la durée de la voyelle tonique de plus de 180 ms.

2.3.2. Au niveau des gestes manuels

Ces disfluences ont ensuite été analysées en interaction avec la tâche (avec gestes *vs*. sans gestes). Nous avons également étudié le taux de gestes répétés d'une part et produits spontanément d'autre part, ainsi que leur type en nous basant sur la typologie de McNeill (1992) :

- Déictiques : gestes référant à quelque chose par le pointage ;
- Iconiques : gestes représentant des mouvements ou formes d'objets ou de personnes ;
- Battements : gestes binaires accompagnant une syllabe ou un phonème, ou encore la parole en rythme.

Nous souhaitons étudier l'impact de plusieurs variables explicatives et leurs interactions, à savoir : âge, NSE, MMSE, groupe (Patient *vs*. Contrôle) et tâche (avec gestes *vs*. sans gestes) sur deux variables distinctes à expliquer, soit le nombre de disfluences au niveau segmental et le nombre de disfluences au niveau suprasegmental. Les participantes étant sollicitées à plusieurs reprises et chaque phrase étant répétée par les 20 participantes, les effets aléatoires 'phrase cible' et 'participant' ont été introduits. Pour le niveau segmental, étant donné que les variables à expliquer sont des variables de comptage et que nous sommes en situation de surdispersion, nous avons opté pour un modèle de régression binomiale négative avec effets aléatoires. Pour le niveau suprasegmental, n'étant pas en situation de surdispersion, le modèle de régression de Poisson avec effets aléatoires a été utilisé et une interaction MMSE*groupe a été ajoutée au vue d'analyses nécessitant des précisions. Pour cela, nous nous sommes servis de la fonction glmmadmb du package glmmADMB du logiciel R. Qui plus est, afin de déterminer si les variables à expliquer sont différentes en fonction du groupe, la fonction glht du package multcomp du logiciel R a été utilisée (Torten et al., 2008).

3 Résultats

3.1 Aux niveaux segmental et suprasegmental

Les participantes atteintes de la MA ont produit significativement plus de disfluences au niveau segmental que les contrôles ($z = 3,82$; $p < 0.0001$) (cf. Figure 2). Par contre, nous n'avons pas trouvé d'effets significatifs ni de l'âge, ni du NSE, ni du MMSE. Une analyse plus fine des erreurs montre que le taux[1] d'erreurs phonétiques des patientes est de 0,82 % et d'erreurs lexicales de 7,69 %. En comparaison le taux d'erreurs phonétiques des contrôles est de 0,48 % et d'erreurs lexicales de 1,02 %. L'analyse statistique a montré que le groupe Patient n'a pas produit significativement plus d'erreurs phonétiques que le groupe Contrôle. Par contre un effet significatif du MMSE a été trouvé ($z = -3,62$; $p = 0,0003$) : plus le MMSE est faible, plus le nombre d'erreurs phonétiques est élevé. En ce qui concerne les erreurs lexicales, les patientes en ont produit significativement plus que les contrôles ($z = 3,99$; $p = 0,0001$).

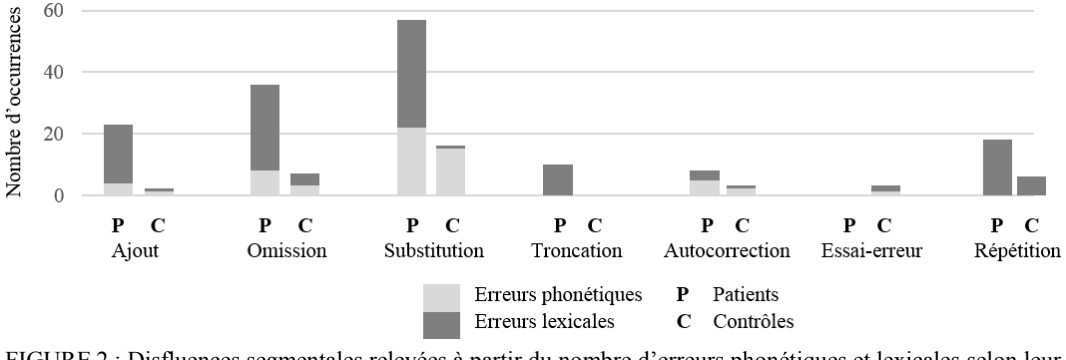
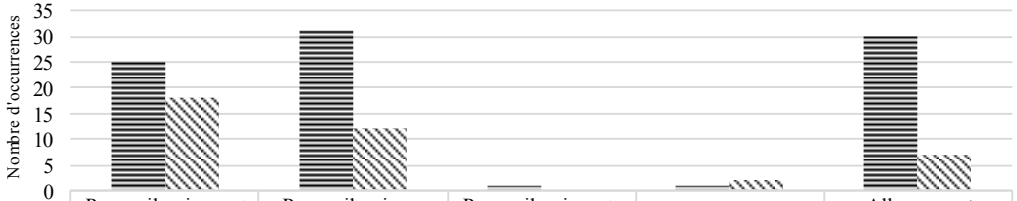


FIGURE 2 : Disfluences segmentales relevées à partir du nombre d'erreurs phonétiques et lexicales selon leur type en fonction du groupe Patient *vs*. Contrôle.

Les participantes atteintes de la MA ont produit significativement plus de disfluences au niveau suprasegmental que les contrôles et d'autant plus que leur MMSE était faible ($z = 3,84$; $p < 0,0004$)

[1] Taux d'erreurs = (valeur théorique – valeur expérimentale) x 100 / valeur théorique

(cf. Figure 3). Le modèle statistique a également montré l'effet de l'âge ($z = -4,45$; $p < 0,01$), mais pas du NSE. Nous avons pu remarquer que 31,43 % des allongements vocaliques produits par les patientes précédaient une pause silencieuse. Concernant l'ensemble des pauses silencieuses, nous avons observé que les patientes ont produit 40,36 % d'entre elles conjointement à des disfluences au niveau segmental, contre 23,33 % pour les contrôles.



	Pause silencieuse < 200 ms	Pause silencieuse [200 ms,1000ms]	Pause silencieuse > 1000 ms	Pause sonore	Allongement vocalique
Patients	25	31	1	1	30
Contrôles	18	12	0	2	7

FIGURE 3 : Disfluences suprasegmentales observées à partir du nombre de pauses et d'allongements vocaliques en fonction du groupe Patient *vs*. Contrôle.

3.2 Au niveau des gestes manuels

Dans la tâche avec gestes imposés, le taux d'erreurs des gestes répétés est de 10 % pour le groupe Patient et de 0,83 % pour le groupe Contrôle. Une analyse plus fine des types de gestes répétés a montré que le taux d'erreurs pour les déictiques est de 15 % pour les patientes et de 0 % pour les contrôles, et pour les iconiques de 10 % pour les patientes et de 1,25 % pour les contrôles. Seul un geste de battement a été produit spontanément par une patiente en lieu et place d'un iconique. Dans cette tâche, nous n'avons pas observé de différence de fluence selon si les gestes étaient ou non répétés.

Pour la tâche sans gestes imposés, les participantes, indépendamment de leur groupe, ont produit spontanément des gestes représentationnels (1 déictique et 4 iconiques) et non-représentationnels (3 battements). Les gestes étaient produits spontanément aussi bien en situation de disfluences qu'en situation de parole fluente.

Un effet significatif de la tâche sur les erreurs phonétiques a été mis en évidence pour les deux groupes avec davantage d'erreurs produites dans la tâche avec gestes imposés ($z = -2,08$; $p = 0,03$). Le modèle statistique montre que l'évolution du taux de disfluences au niveau segmental en fonction de la valeur du MMSE est significativement différente entre les patientes et les contrôles dans la tâche avec gestes ($p < 0.01$) par rapport à la tâche sans gestes ($p < 0.01$). Dans le groupe Patient, un effet significatif de la tâche a été observé ($z = 2,85$; $p = 0,02$). Dans la tâche avec gestes, plus le MMSE est élevé, plus les patientes produisent des disfluences au niveau suprasegmental, alors que dans la tâche sans gestes, plus le MMSE est élevé, moins les patientes produisent de disfluences. En comparaison, pour le groupe Contrôle, plus le MMSE est élevé, plus le taux de disfluences diminue rapidement dans la tâche sans gestes par rapport à la tâche avec gestes ($z = 2,85$; $p = 0,02$).

4 Discussion

Cette étude permet de confirmer une partie de notre hypothèse de départ : les personnes atteintes de la MA ont effectivement une atteinte significative de la fluence qui se traduit dans les niveaux segmental et suprasegmental par une baisse en comparaison aux productions des personnes appariées en âge et en sexe ne souffrant pas de TNC, conjointe à un taux de gestes manuels répétés moins important pour le groupe Patient que pour le groupe Contrôle. Néanmoins, nous n'avons pas observé de différence entre les deux groupes au niveau du taux de gestes manuels produits

spontanément. Nous avons également trouvé un impact de la répétition de gestes manuels sur la fluence aux niveaux segmental et suprasegmental, mais indépendamment du groupe.

Nos résultats au niveau segmental vont dans le sens des études qui attribuent l'augmentation des disfluences à un trouble de la récupération lexico-sémantique dans la MA et plus précisément à une aphasie (ex. : Tran et al., 2011). L'analyse des types d'erreurs montre un nombre plus important de substitutions et d'omissions produites par les patientes, ce qui est en accord avec la littérature qui décrit la paraphasie et l'anomie comme premiers symptômes de l'aphasie. Qui plus est, seules les patientes ont produit des troncations, toutes au début des énoncés, ce qui va dans le sens d'un effet de récence et suggère un empan mnésique réduit chez les personnes atteintes de la MA dus à des troubles de la mémoire de travail typique de cette maladie (Baddeley, 1992 ; Amieva et al., 2014). Concernant les erreurs phonétiques, nos résultats ne montrent pas de différence significative entre les contrôles et les patientes, même si le taux d'erreurs phonétiques augmente avec des TNC plus importants, ce qui est cohérent avec les travaux cités dans Aubin & Le Gall (2003) qui stipulent que la praxie est plutôt bien préservée aux premiers stades de la maladie.

Au niveau suprasegmental, de même que montré par Lee (2012), la parole des personnes atteintes de la MA se caractérise par une diminution de la fluence caractérisée par des hésitations – telles que des allongements vocaliques plus fréquents et des pauses silencieuses plus fréquentes et plus longues – par rapport à des sujets contrôles.

Contrairement aux précédents travaux sur les gestes manuels des personnes atteintes de la MA (Schiaratura et al., 2015 ; Hubbard et al., 2002), nous avons trouvé une diminution du taux de gestes répétés par rapport aux contrôles qui peut être expliquée par le fait que dans ces études la tâche consistait en une tâche de parole spontanée et non d'une tâche de répétition comme dans notre étude. Cette diminution peut être due à un phénomène de double tâche que Baddeley explique par une atteinte du 'buffer épisodique' contrôlé par l'administrateur central de la mémoire de travail (Amieva et al., 2014). Un impact de la répétition de gestes sur le taux d'erreurs phonétiques a également été mis en exergue aussi bien chez les patientes que les contrôles. Ce phénomène moteur pourrait être un argument en faveur de la co-expressivité entre gestes manuels et parole qui impliquerait un lien articulatoire et pas uniquement sémiotique comme avancé par McNeill (1992). Un impact de la répétition de gestes sur le taux de disfluences au niveau suprasegmental a également été montré. Le taux plus important de disfluences au niveau suprasegmental avec un MMSE plus élevé pourrait être expliqué par un effet de double tâche dans le groupe Contrôle et pour les patientes avec une atteinte cognitive légère (Baddeley, 1992) et par une anosognosie chez les patientes avec une atteinte cognitive légère, c'est-à-dire par une perte du malade de sa propre connaissance ou de sa propre conscience plus importante avec l'avancée des TNC (Amieva et al., 2014). Dès lors, ces résultats interrogent le fait qu'une tâche de répétition avec *vs*. sans gestes imposés permet de mesurer l'impact des gestes manuels ou plutôt la charge cognitive des participants.

Il est à noter que nos résultats n'ont pas trouvé d'effet significatif du NSE ni pour le groupe Contrôle ni pour le groupe Patient en interaction avec le MMSE, contrairement à la littérature (Lee, 2012 ; Stern et al., 1994).

Ces résultats préliminaires nécessitent une étude plus approfondie sur une population plus large afin d'éviter, voire au moins de minimiser, les effets de la variabilité interindividuelle. Cette étude est toujours en cours, notamment auprès de personnes atteintes de la MA à un stade sévère afin de continuer l'investigation des troubles de la communication multimodale des personnes atteintes de cette maladie.

Remerciements

Cette recherche a été financé par le Ministère français de l'enseignement supérieur et de la recherche, la Fondation SFR Santé et Société et le Pôle Grenoble Cognition d'avoir financé la réalisation de ce projet. Merci au CH de Tullins et au MAPAD Arc-en-ciel pour leur aide quant au recrutement des participants, ainsi qu'aux participantes et leurs familles.

Références

ALBERT M.L. (2011). *Neuropsychologie du vieillissement normal*. Boston: University Press, 79-86.

AMIEVA H., BELLIARD S. & SALMON E. (éds.). (2014). *Les démences. Aspects cliniques, neuropsychologiques, physiopathologiques et thérapeutiques*. Louvain-la-Neuve: De Boeck Supérieur.

AUBIN G. & LE GALL D. (2003). *L'apraxie*. Paris: Solal.

BADDELEY A. (1992). *La mémoire humaine : théorie et pratique*. Grenoble: Presses universitaires.

BEESON P.M., HOLLAND A.L. & MURRAY L.L. (1997). Naming famous people: an examination of tip-of-the-tongue phenomena in aphasia and Alzheimer's disease. *Aphasiology* 11, 323-336.

CAUSSADE D., GAUBERT F., SERIEUX M., HENRICH-BERNARDONI N., COLLETTA J.-M. & VALLEE N. (2015). Hand gestures and speech impairments in spoken and sung modalities in people with Alzheimer's disease. *Actes de conférence de GESPIN,* Nantes.

DE RUITER J.P. (2009). Can gesticulation help aphasic people speak, or rather, communicate? *International Journal of Speech-Language Pathology*, 124-127.

FEYEREISEN P., BERREWAERTS J. & HUPET M. (2007). Pragmatic skills in the early stages of Alzheimer's disease: an analysis by means of a referential communication task. *Int J Lang Com Dis* 42, 1-17.

FOLSTEIN M.F., FOLSTEIN S.E. & MCHUGH P.R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* 12 (3), 189–198.

FÜLLGRABE C. & MOORE B.J.C. (2014). Effects of age and hearing loss on stream segregation based on interaural time differences. *The Journal of the Acoustical Society of America* 136(2), 185-191.

GAYRAUD F., LEE H.R., HIRSCH F. & BARKAT-DEFRADAS M. (2011). Perturbations phonologiques et maladie d'Alzheimer : la fin d'un mythe ? *Actes de conférence des 4èmes Journées de Phonétique Clinique,* Strasbourg.

HUBBARD G., COOK A., TESTER S. & DOWNS M. (2002). Beyond words: older people with dementia using and interpreting non-verbal behavior. *Journal of Aging Studies* 16, 155-167.

JOUBERT A. & LE ROUZO M.L. (2000). A comparison of the tip of the tongue phenomenon between elderly and young people. *Second International Conference on the Mental Lexicon,* Montréal.

KALAFAT M., HUGONOT-DIENER L. & POITRENAUD J. (2003). Standardisation et étalonnage français du 'Mini Mental State' (MMS) version GRECO. *Revue de Neuropsychologie* 13 (2), 209-236.

LEE H. (2012). *Langage et maladie d'Alzheimer : analyse multidimensionnelle d'un discours pathologique*. Thèse en Sciences du langage, Université Paul-Valéry Montpellier 3.

LUCHESI CERA M.L., ORTIZ K.Z., FERREIRA BERTOLUCCI P.H. & CIANCIARULLO MINETT T.S. (2013). Speech and orofacial apraxias in Alzheimer's disease. *International Psychogeriatrics* 25(10), 1679-1685.

MCNEILL D. (1992). *Hands and mind. What gestures reveal about thought*. Chicago and London: The University of Chicago Press.

MILLER G.A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review* 63(2), 81-97.

MORSELLA E. & KRAUSS R.M. (2004). The Role of Gestures in Spatial Working Memory and Speech. *The American Journal of Psychology* 117, 411-424.

PENNY L., MITCHELL S., SAUNDERS N., HUNWICK J., MITCHARD H., & VRLIC M. (1996). Some aspects of speech and voice in healthy ageing people. Actes de la *Sixth Australian International Conference on Speech Science and Technology*, Adelaïde, SA.

SCHIARATURA L.T. (2015). Expression verbale et gestualité dans la maladie d'Alzheimer : une étude en situation d'interaction sociale. *Geriatr Psychol Neuropsychiatr Vieil* 13(1), 97-105.

SCHÖTZ S. (2006). *Perception, Analysis and Synthesis of Speaker Age*. Lund: Media-Tryck.

STERN Y., GURLAND B., TATEMICHI T. K., TANG M. X., WILDER D., & MAYEUX R. (1994). Influence of education and occupation on the incidence of Alzheimer's disease. *Journal of American Medical Association* 271, 1004-1010.

TELLIER M. (2009). The development of gesture. *Language development over the lifespan*, 191-216.

TORTEN H., BERTZ F. & WESTFALL P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal* 50(3), 346-363.

TRAN T.M. & GODEFROY O. (2011). La Batterie d'Evaluation des Troubles Lexicaux : effets des variables démographiques, linguistiques, reproductibilité et normes. *Revue de Neuropsychologie* 3 (1), 52-69.

Disfluences normales *vs*. Disfluences sévères : une étude acoustique

Ivana Didirkova[1], Camille Fauth[2], Fabrice Hirsch[1], Giancarlo Luxardo[1], Sascha Diwersy[1]

(1) Université Paul Valéry - Montpellier III, Praxiling UMR 5267, CNRS
(2) Université de Strasbourg, Institut de Phonétique de Strasbourg – IPS & U.R. 1339 Linguistique, Langues et Parole – LilPa, E.R. Parole et Cognition

ivana.didirkova@univ-montp3.fr, cfauth@unistra.fr, fabrice.hirsch@univ-montp3.fr, giancarlo.luxardo@univ-montp3.fr, sascha.diwersy@univ-montp3.fr

RESUME

L'objectif de cette recherche est d'étudier les caractéristiques acoustiques et perceptives des disfluences normales et sévères. Pour ce faire, un jury d'auditeurs experts a relevé les disfluences sévères et normales de 4 locuteurs qui bégaient ainsi que les accidents de parole de 4 sujets normo-fluents. Une analyse acoustique portant sur des paramètres tels que la durée de la disfluence, le nombre de disfluences ou encore sur la présence d'éléments prosodiques particuliers a été menée sur les seules disfluences ayant été relevés par l'ensemble du jury. Nos résultats montrent que si les prolongations et les répétitions sont bien évidemment catégorisées comme sévères respectivement en fonction de leur durée et du nombre d'éléments réitérés, d'autres paramètres sont également significatifs, tels que la présence ou non d'une tension audible, le type d'éléments répétés ou encore le fait que la syllabe soit ou non clivée.

ABSTRACT

Normal disfluences vs. stuttering-like disfluences: an acoustic study

The aim of this research is to study acoustic and perceptive characteristics of stuttering-like disfluencies and other disfluencies. An acoustic analysis has been conducted on stuttering-like disfluencies produced by four persons who stutter and on other disfluencies produced by four non-stutterers that were annotated by all members of a jury of expert listeners. Different parameters, including duration and number of disfluencies or their prosodic particularities, were studied. Our results show that, in addition to the duration and number of disfluencies, other parameters can be used to differentiate other and stuttering-like disfluencies, such as presence or not of audible tensions within the disfluency, type of repeated elements (e.g. word vs. sound) or location of the disfluency.

MOTS-CLÉS : bégaiement, disfluences, bégayages

KEYWORDS: stuttering, disfluencies, stuttering-like disfluencies

1 Introduction

Les accidents de parole (ou disfluences) jalonnent les productions orales, surtout lorsque celles-ci ne sont pas préparées à l'avance (Corley & Stewart, 2008). La fréquence d'apparition ainsi que la nature de ces disfluences dépendent de nombreux critères et il arrive parfois que la sévérité de ces accidents de parole entrave l'intelligibilité de l'énoncé (MacGregor *et al.*, 2009). C'est d'autant plus le cas pour les pathologies affectant le rythme et la fluence comme le bégaiement qui se caractérise notamment par des répétitions, des prolongations, des blocages, etc. (Van Riper, 1973; Monfrais-Pfauwadel, 2014).

1.1 Les disfluences « normales »

Plusieurs recherches ont été réalisées sur les accidents présents dans la parole de locuteurs normo-fluents pour évaluer leur importance quantitative et leurs caractéristiques qualitatives.

C'est le cas d'une étude menée par Lutz et Mallard (1986), qui portait sur les disfluences chez des jeunes adultes normo-fluents et qui relève que 3,2% et 3,4% de mots sont disfluents respectivement chez des sujets masculins et des sujets féminins, les syllabes touchées représentant 2,4% et 2,7% des syllabes prononcées. Les catégories de disfluences les plus représentées durant la tâche de conversation sont celles des interjections (27,3%), des révisions (25%) et des répétitions de mots et / ou de phrases (15,2%). Inversement, les sujets ayant pris part à cette étude n'ont produit aucune répétition de partie de mot. De plus, les prolongations, les phonations arythmiques et les phrases incomplètes ont été très peu présentes (moins de 0,1%). Ces résultats sont confortés par ceux obtenus par Duchin et Mysak (1987) ou encore Roberts, Meltzer et Wilding (2009) qui ont également observé davantage d'interjections, de révisions et de répétitions de mots. Teixeira *et al.* (2012) ont, quant à eux, observé que le pourcentage de silence marqué durant une tâche de conversation était compris entre 11% du temps de parole chez les sujets les moins disfluents et 26,30% pour les locuteurs présentant le plus d'accidents de parole.

Quant aux propriétés acoustiques des disfluences normales, conformément à Levelt et Cutler (1983), une emphase contrastive peut être observée sur 53% de corrections produites suite à une erreur, tandis que seules 19% de corrections faisant suite à un énoncé inapproprié subissent le même « marquage ». Cole *et al.* (2005) ont observé des contours de *f0* similaires en phase de reparandum et d'altération (McTear, 2004). En d'autres termes, la prosodie de l'item à corriger et celle de la version corrigée seraient comparables. Shriberg (1995) s'est intéressée aux répétitions dites perspectives et rétrospectives. Les premières, précédées et suivies d'une pause, serviraient à garder un contact avec l'interlocuteur durant la disfluence. Les répétitions rétrospectives, précédées mais non suivies d'une pause, seraient simplement utilisées pour marquer la continuité avec ce qui a été dit précédemment. Se basant sur cette distinction, Shriberg (1995) observe que les répétitions rétrospectives subissent un allongement de la première répétition, mais pas de la seconde. En revanche, dans le cas des répétitions prospectives, le deuxième élément est allongé. De même, tandis que les répétitions prospectives ont subi une baisse continue de la fréquence fondamentale tout au long de la disfluence, ce même paramètre atteint approximativement le même niveau au début de chacune des deux répétitions rétrospectives.

1.2 Disfluences sévères

Plusieurs études ont été menées afin d'essayer de définir la limite entre une disfluence normale et une disfluence sévère, retenant divers critères. L'une des études les plus connues à ce sujet, celle de Van Riper (1973), estime par exemple qu'une prolongation d'une durée supérieure à 1 seconde peut être qualifiée comme sévère. D'autres études, comme celle menée par Lechta & Štenclová (2009), ont conclu que le curseur séparant les accidents de parole normaux des bégayages se situait plutôt à 500 ms. De même, toujours selon Van Riper (1973), afin que la disfluence soit perçue comme normale, le nombre de répétitions doit être inférieur à deux. D'autres restent moins prescriptifs, estimant toutefois que plus le nombre de répétitions est élevé et plus il y a de risque que la disfluence soit sévère (Ward, 2006). L'élément disfluent peut également jouer un rôle dans la distinction entre ces deux types de disfluences dans la mesure où les accidents de parole touchant des syllabes ou des phonèmes (Remacle, 2011), ou ceux survenant à l'intérieur d'une syllabe (Zellner, 1992), seront source d'une plus grande perturbation pour la communication.

1.3 Problématique et hypothèses

Deux objectifs ont rythmé ce travail de recherche. Le premier était de vérifier quels indices acoustiques permettaient à des auditeurs de catégoriser des disfluences comme normales ou sévères. Nous pensons en effet que certains indices temporels (durée de la disfluence,…) et prosodiques (intensité, type de disfluences,...) sont utilisés par les auditeurs pour distinguer un simple accident de parole d'un bégayage.

Le second objectif de cet article consiste à savoir si des différences sont observables entre les disfluences normales produites par des personnes qui bégaient et celles produites par des sujets ne présentant aucun trouble du langage. Plus concrètement, il s'agit de vérifier si la parole bègue se distingue de la parole normo-fluente uniquement par rapport aux disfluences catégorisées comme sévères ou si les accidents de parole plus classiques présentent également des dissimilitudes entre les deux groupes. Etant donné que les sujets atteints de bégaiement ont un autocontrôle plus important de leur mouvement (Monfrais-Pfauwadel, 2014), on peut supposer retrouver par exemple des différences de durée lors des pauses silencieuses.

2 Protocole expérimental

2.1 Corpus, sujets

Pour mener à bien notre étude, huit locuteurs ont été enregistrés : quatre personnes qui bégaient (moyenne d'âge = 35,25 ans, ET = 10,532 ; trois de sexe masculin et une de sexe féminin) appariées en sexe et en âge avec quatre sujets normo-fluents (moyenne d'âge = 37,75 ans, ET = 14,43) ont accepté de prendre part à cette recherche. Les sujets avaient pour tâche de se présenter brièvement, de décrire une journée-type et / ou un film qu'ils ont vu récemment.

2.2 Méthode d'analyse

Une fois les entretiens réalisés, une transcription orthographique a été faite sur 2 mn de prise de parole pour chacun des 8 sujets. Celle-ci a ensuite été soumise à un jury de trois experts

phonéticiens. Ces derniers avaient pour consigne d'écouter les 2x8 mn de temps d'enregistrements et de repérer sur la transcription, l'ensemble des disfluences présentes dans la parole des locuteurs qui bégaient et des sujets normo-fluents. Les membres du jury avaient également pour tâche de distinguer les disfluences sévères de celles qui leur paraissaient normales : les premières citées se définissaient comme celles passant inaperçues dans le discours tandis que les secondes sont celles qui pourraient laisser penser que le sujet parlant est bègue. Signalons que seules les disfluences ayant été relevées et catégorisées de la même façon (en termes de sévérité) par tous les membres du jury ont été retenues pour la suite de l'étude.

Ce travail achevé, une analyse fine des disfluences retenues a été effectuée en tenant compte des différentes caractéristiques généralement admises dans la littérature pour les décrire ; ces traits sont résumés par Monfrais-Pfauwadel (2014). Il s'agissait de noter :

- Le type de disfluences ;
- La durée totale de la disfluence ;
- La présence ou non d'un clivage de syllabe ;
- La présence ou non d'une tension audible due à une variation de l'intensité ;
- La localisation de la disfluence (inter-phrases, inter-propositions, inter-syntagmes, intra-syntagme).

De même, des paramètres plus spécifiques ont également été étudiés en fonction du type de disfluence comme :

- Le nombre de répétitions ;
- Le type de répétitions (phrase, segment de phrase, mot, mot monosyllabique, phonème) ;
- La durée des blocages et des pauses d'hésitation ;
- La durée des prolongations ;
- Le type d'élément prolongé.

Au total, 323 disfluences ont été analysées : 73 disfluences sévères et 121 disfluences normales produites par les personnes qui bégaient et 129 accidents de parole présents dans la parole des sujets normaux fluents. Des analyses statistiques ont été effectuées (1) pour vérifier si chacun des paramètres mesurés permettait de différencier les disfluences sévères des accidents de parole présents dans la parole normo-fluente et (2) pour savoir si les traits étudiés rendaient possible la distinction entre les disfluences normales produites par les personnes bègues et celles des sujets de contrôle.

3 Résultats

3.1. Relevé des disfluences

L'étude de la Figure 1 révèle que les fillers sont les disfluences les plus présentes dans le discours des locuteurs normo-fluents ayant pris part à cette étude puisqu'ils représentent 40% des accidents de parole. En outre, on relève que les pauses d'hésitation sont également très fréquentes dans la mesure où elles constituent 32% des disfluences. Signalons aussi que les allongements de sons semblent significatifs, étant donné qu'ils représentent 19% des disfluences relevées chez les locuteurs ne présentant pas de trouble d'élocution.

Ces mêmes types de disfluences sont présents dans la parole des locuteurs qui bégaient mais à un degré moindre, comme en atteste la Figure 2. En effet, les fillers constituent 24% des accidents de parole produits pour cette catégorie de locuteurs. Quant aux pauses d'hésitation et aux prolongations

dites « normales », elles constituent respectivement 14% et 12%. A ces disfluences normales, s'ajoutent un certain nombre de disfluences plus sévères tels que les blocages, qui constituent 19% des accidents de parole, les répétitions sévères, qui représentent 13% des disfluences et les prolongations considérées comme anormales (9% des disfluences).

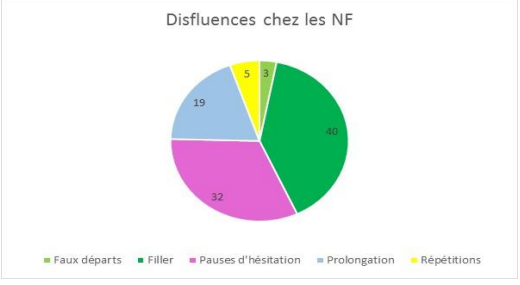


Figure 1 : Disfluences produites par les sujets normo-fluents (en pourcentages).

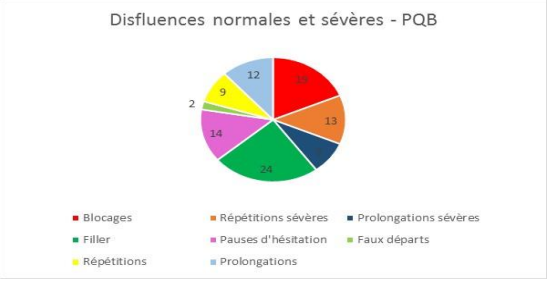


Figure 2 : Disfluences produites par les personnes qui bégaient (en pourcentages).

En résumé, la parole bègue se définit par la présence de disfluences classiques, également présentes dans la parole des sujets normo-fluents, mais également par des accidents de parole plus sévères qui caractérisent le bégaiement. En outre, on remarque également un certain nombre de disfluences, tels que les prolongations et les répétitions qui peuvent être considérées comme normales ou sévères. La question qui se pose alors est de savoir où se situe la séparation entre ces catégories de disfluences qui peuvent aussi bien être classées dans les non-pathologiques que dans les pathologiques.

3.2. Prolongations normales *vs*. Prolongations pathologiques

L'étude de la durée des prolongations normales produites par les sujets normo-fluents en comparaison avec celle les disfluences sévères présentes dans la parole bègue s'est révélée significative (p=0,01431). La Figure 3 montre que les prolongations de sons perçues comme normales durent 526 ms en moyenne. Quant aux prolongations jugées comme sévères, elles sont de 837 ms. On relèvera également qu'en plus d'être plus longue, la durée des allongements pathologiques présente une variabilité plus grande, l'écart-type étant de 593 ms pour ce type de disfluences alors qu'il n'est que de 209 ms pour les prolongations non-pathologiques.

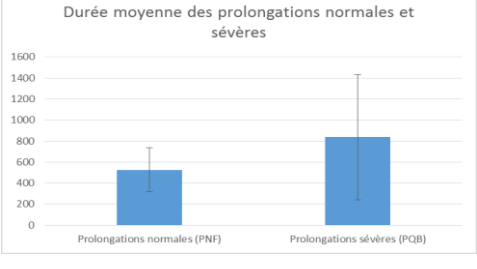


Figure 3 : durée moyenne des prolongations normales produites par les locuteurs normo-fluents et les prolongations sévères présentes dans la parole des sujets qui bégaient.

En plus de la durée, d'autres paramètres sont également à prendre en considération en vue de distinguer les prolongations normales des prolongations sévères. C'est le cas de la présence ou non

d'une tension audible, se caractérisant notamment par une augmentation de l'intensité, qui accompagne généralement les prolongations sévères (p=0,00021). En outre, si les allongements normaux ont lieu essentiellement sur des voyelles, celles considérés comme sévères portent aussi bien sur des éléments consonantiques que vocaliques (p=0,00047).

Signalons enfin que la comparaison entre les allongements normaux produits par les sujets normo-fluents et ceux des locuteurs qui bégaient n'est statistiquement pas significative.

3.3. Répétitions normales *vs.* Répétitions pathologiques

L'étude de la durée des répétitions ne s'est pas révélée pertinente (p>0,05). En revanche, le nombre de répétitions est un critère qui s'est montré significatif (p=0,049).

La Figure 4 présente ainsi le nombre moyen de répétitions en cas de disfluences normales et pathologiques. On constate qu'il y a 1,4 éléments répétés lors des disfluences non-pathologiques (Ecart-type : 0,534), tandis que 2,09 (ET : 0,97) segments sont réitérés lors des bégayages.

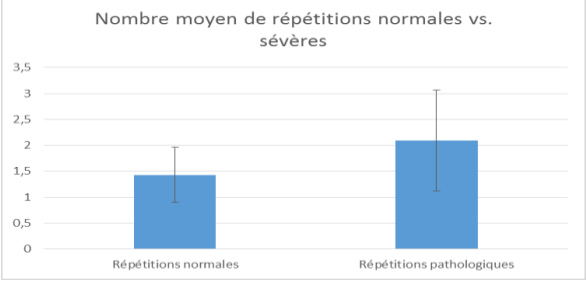
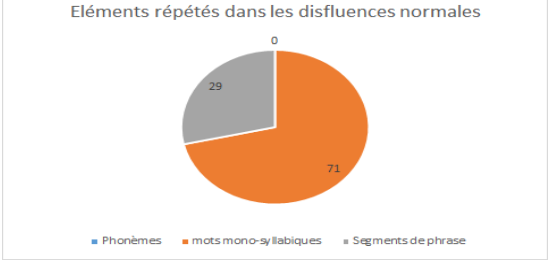


Figure 4 : nombre moyen de répétitions lors des disfluences catégorisées comme normales (à gauche) et comme sévères (à droite).

D'autres éléments entrent également en jeu pour classifier des répétitions comme pathologiques ou non. C'est le cas de la présence ou non d'une tension audible : les répétitions considérées comme sévères sont significativement accompagnées d'un effort excessif se caractérisant par une montée de l'intensité (p=0,02781). De même, nous avons constaté que les éléments répétés n'étaient pas systématiquement les mêmes dans les disfluences normales et sévères : les répétitions de phonème sont davantage caractérisées comme des disfluences sévères (voir Figures 5 et 6).



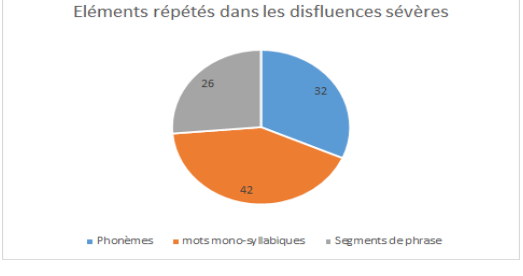


Figure 5 : type d'éléments répétés dans les

Figure 6 : type d'éléments répétés dans les

disfluences normales (en pourcentages). disfluences sévères (en pourcentages).

Signalons que la comparaison des disfluences produites par les sujets normo-fluents et les personnes qui bégaient n'a pas révélé de différences significatives, ni pour la durée des répétitions, ni pour leur nombre. Enfin, aucune tension n'a été repérée dans les disfluences normales produites par les personnes atteintes de ce trouble du rythme de la parole.

3.4. Blocages silencieux et pauses d'hésitation

Dans cette partie, nous avons fait le choix de comparer les blocages silencieux aux pauses d'hésitation dans la mesure où ces deux événements présents sur le signal de parole se caractérisent par une interruption du signal sonore à un instant non choisi par le locuteur. Signalons que nous parlons de pauses d'hésitation pour désigner les interruptions du signal sonore qui sont précédées d'une pause pleine de type « euh ». Par la suite, nous avons voulu vérifier si des différences apparaissaient entre les pauses d'hésitation produites par les locuteurs de contrôle et celles des personnes qui bégaient.

La confrontation des durées des blocages silencieux avec les pauses d'hésitation n'est pas significative (p=0,142). Par conséquent, le jury d'auditeurs a dû fonder sa classification sur d'autres paramètres. On notera ainsi que le clivage de la syllabe a pu constituer un critère de différenciation entre ces deux types d'accidents de parole : les blocages silencieux « cassent » régulièrement la syllabe dans laquelle ils apparaissent, ce qui n'est pas le cas des pauses d'hésitation qui préservent l'intégrité des syllabes (p=0,000). Par ailleurs, nous avons également constaté que le blocage silencieux était généralement suivi ou précédé d'un effort audible se caractérisant par une montée de l'intensité (p=0,000).

Quant à la comparaison de la durée des pauses d'hésitation produites par les personnes normo-fluentes avec celles des sujets bègues, elle s'est révélée significative (p=0,02116). La Figure 7 montre que les interruptions dues à une hésitation durent 1098 ms (ET : 522) en moyenne chez les locuteurs de contrôle alors qu'elles mesurent 680 ms (ET : 291) chez les personnes bègues. Signalons cependant que ce résultat doit être pris avec précaution dans la mesure où le débit de parole, différent d'un sujet à un autre peut venir expliquer ces données.

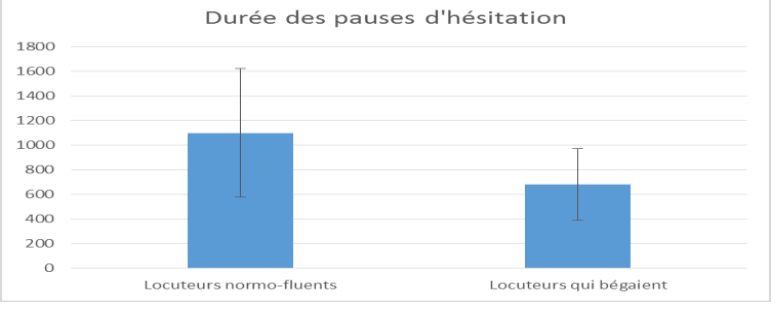


Figure 7 : Durée des pauses d'hésitation chez les locuteurs normo-fluents et les sujets qui bégaient.

4 Synthèse et perspectives

Les résultats obtenus à travers cette étude pilote ont permis de rappeler que la parole bègue ne se caractérise pas uniquement par des disfluences sévères : celle-ci contient à la fois des accidents de parole classiques, que l'on retrouve également dans la parole des sujets normo-fluents, et des disfluences d'une sévérité accrue, qui sont typiques de la parole bègue.

L'étude de la perception des disfluences associées au bégaiement par notre jury d'expert a également révélé que plusieurs facteurs expliquent leur choix. Si les prolongations sévères sont, comme attendu, plus longues que celles catégorisées comme normales, on relèvera qu'elles sont régulièrement accompagnées d'une tension qui est audible et qu'elles se manifestent sur des segments qui ne sont généralement pas prolongés dans le cas des disfluences normales. Quant aux disfluences se caractérisant par des répétitions, on notera d'une part que le nombre d'éléments réitérés est plus important en cas d'accidents de parole qualifiés de sévères et, d'autre part, que le type d'éléments bégayés n'est pas systématiquement le même : les disfluences sévères peuvent se caractériser par des répétions de phonème, ce qui n'est pas le cas dans les accidents de parole plus classiques. Enfin, nous avons également pris le parti de comparer les blocages silencieux aux pauses d'hésitation dans la mesure où ces deux faits de parole se définissent par une interruption du signal acoustique réalisée à un instant non voulu par le locuteur. Cette comparaison a révélé que la durée n'était pas un paramètre significatif pour distinguer ces deux types d'interruption du signal sonore. En revanche, la localisation de cette interruption semble jouer un rôle majeur dans la catégorisation de la disfluence puisqu'on remarquera que seules les disfluences sévères peuvent se définir par un arrêt du signal sonore à l'intérieur de la syllabe. De même, des tensions sont très souvent audibles avant le blocage sur le signal de parole.

Quant à la comparaison des disfluences normales produites par les sujets bègues avec celles des locuteurs non-bègues, elle ne montre pas de résultats significatifs, à l'exception de la durée des pauses d'hésitation plus courtes chez le second groupe de locuteurs cité. Cependant, ce résultat est à prendre avec précaution et devrait être conforté à l'aide d'autres analyses sur le sujet.

Comme autre perspective à ce travail, nous pensons qu'il serait intéressant d'augmenter le nombre de locuteurs normo-fluents et bègues afin d'amplifier le nombre de disfluences normales et sévères à étudier. De même, une étude articulatoire devrait également être menée en parallèle à cette recherche afin de trouver les éléments moteurs à l'origine de ces différences en termes de qualité acoustique entre disfluences normales et sévères. Enfin, il serait également souhaitable de prolonger cette recherche en étudiant les disfluences normales qui n'ont pas été abordées dans cet article (fillers, faux départs, ajouts de mots d'appui,…) afin d'observer comment les personnes qui bégaient utilisent ces éléments pour masquer leur trouble d'élocution.

Remerciements

Ce travail a été financé par un projet l'IDEX (Initiative d'excellence) 2015 – Attractivité «Arythmique : un logiciel de diagnostic et de suivi automatique de patients atteints de troubles de la parole et du langage liés au rythme» porté par Madame Camille FAUTH.

Références

COLE J., HASEGAWA-JOHNSON M., SHIH CH., KIM H., LEE E.-K., LU H., MO, Y., YOON T.-J. (2005). Prosodic parallelism as a cue tu repetition disfluency. Actes de *Disfluency in Spontaneous Speech Workshop*. http://prosody.beckman.illinois.edu/pubs/15_Cole_etal_2005_DiSS05.pdf

CORLEY M., STEWART O.W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, 589-602.

DUCHIN S.W., MYSAK E.D. (1987). Disfluency and rate characteristics of young adult, middle-aged, and older males. *Journal of Communication Disorders*, 20(3), 245-257.

GOLDMAN-EISLER F. (1968). *Psycholinguistics: Experiments in Spontaneous Speech*. London: Academic Press.

LECHTA, V., ŠTENCLOVÁ, Ľ. (2009). *Zajakavé dieťa doma a v škole?* Bratislava: SZU.

LEVELT W.J.M., CUTLER A. (1983). Prosodic Marking in Speech Repair. *Journal of Semantics*, 2(2), 205-218.

LUTZ K.C., MALLARD A.R. (1986). Disfluencies and rate of speech in young adult nonstutterers. *Journal of Fluency Disorders*, 11(4), 307-316.

MAC GREGOR L.J., CORLEY M., DONALDSON D.I. (2009). Not all disfluencies are equal: The effects of disfluent repetitions on language comprehension. *Brain and language*, 11, 36-45.

MCTEAR M.F. (2014). *Spoken Dialogue Technology: Toward the Conversational User Interface*. London: Springer Science & Business Media.

MONFRAIS-PFAUWADEL M.-C. (2014). *Bégaiement, bégaiements : Un manuel clinique et thérapeutique*. Paris : De Boeck-Solal.

REMACLE M. (2011). Les symptômes phoniatriques dans le bégaiement. In : *Les bégaiements de l'adulte*, Wavre : Mardaga, 189-197.

ROBERTS P.M., MELTZER A., WILDING J. (2009). Disfluencies in non-stuttering adults across sample lengths and topics. *Journal of Communication Disorders*, 42(6), 414-427.

SHRIBERG E. (1995). Acoustic properties of disfluent repetitions. Actes de *International Congress of Phonetic Sciences*, 384-387.

TEIXEIRA J.P., FERNANDES, M.G., COSTA, R.A. (2012). MEASURE AND COMPARISON OF SPEECH PAUSE DURATION in Subjects with Disfluency Speech. *Procedia Technology*, 5, 812-819.

VAN RIPER CH. (1973). *The treatment of stuttering*. Englewood Cliffs, NJ: Prentice-Hall.

ZELLNER B. (1992). Le bé-bégayage et euh…., l'hésitation en français spontané. Actes des *19ème JEP*, 481-487.

WARD D. (2006). *Stuttering and cluttering: frameworks for understanding and treatment*. New York: Psychology Press.

La distinction entre les paraphasies phonétiques et phonologiques dans l'aphasie : Etude de cas de deux patients aphasiques

Clémence Verhaegen[1], Véronique Delvaux[1,2], Kathy Huet[1], Sophie Fagniart[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) Fond National de la Recherche Scientifique, Belgique
clemence.verhaegen@umons.ac.be

RESUME

La spécificité phonologique ou phonétique des erreurs de production orale observées chez les patients aphasiques reste débattue. Cependant, la distinction entre ces deux types d'erreurs est fréquemment basée sur des analyses perceptives qui peuvent être influencées par le système perceptif de l'expérimentateur. Afin de pallier ce biais, nous avons réalisé des analyses acoustiques des productions de deux patients aphasiques, dans une tâche de répétition de non-mots. Nous nous sommes centrés sur l'analyse de consonnes occlusives. Les résultats ont montré la présence de difficultés de gestion du voisement chez les deux patients, indiquant la présence de troubles phonétiques. En outre, les résultats montrent une grande diversité des manifestations des troubles langagiers des patients ainsi que l'intervention potentielle de stratégies de compensation de leurs difficultés. L'intérêt de procéder à des analyses acoustiques précises utilisant des indices multiples est discuté.

ABSTRACT

The distinction between phonetic and phonological paraphasias in aphasia: A multiple case-study of aphasic patients.

The phonological or phonetic specificity of language production errors in aphasic patients is still debated. However, this distinction is often based on perceptual analyses of the patient's errors, that could be influenced by the experimenter's perceptual system. In our study, in order to clearly attribute these errors affecting phonemes to phonetic difficulties when appropriate, we conducted acoustic analyses of the language production errors of two aphasic patients, on a nonword repetition task. We focused on the analysis of stop consonants. Our results showed voicing difficulties in both patients which suggests phonetic impairment. Moreover, we also found great diversity in the manifestations of our patients' difficulties as well as the potential intervention of palliative strategies used by the patients in order to compensate for their impairment. The interest of acoustic analysis of the aphasic patients' productions using multiple indexes is discussed.

MOTS-CLES : erreur phonétique, erreur phonologique, aphasie, analyse acoustique, délai d'établissement du voisement

KEYWORDS: phonetic errors, phonological errors, aphasia, acoustic analysis, voice onset time

1 Introduction

En aphasiologie, une question importante est de déterminer le niveau de représentations langagières altéré qui est à l'origine des erreurs de production du langage chez les patients. Il est généralement admis que les erreurs peuvent être dues à une atteinte au niveau soit de la sélection des informations lexico-sémantiques relatives au mot cible, soit de la sélection de la forme phonologique abstraite du mot et de la planification des unités sublexicales qui le constituent, voire au plan phonétique, où le locuteur sélectionne et exécute les programmes moteurs articulatoires nécessaires à l'articulation du mot cible (e.g., Levelt, Roelofs & Meyer, 1999). Dans la littérature, deux types d'erreurs ont été attribuées à un trouble au niveau des processus phonologiques ou de la programmation motrice : les paraphasies phonologiques et les paraphasies phonétiques. Les paraphasies phonologiques consisteraient en substitutions, transpositions, ajouts ou suppressions de phonèmes tandis que les paraphasies phonétiques se manifesteraient par des distorsions des réalisations par rapport à l'attente normative (Laganaro, 2015). Cependant, la distinction entre ces deux types de paraphasies n'est pas toujours aisée et est généralement basée sur des analyses perceptives des erreurs du patient (e.g., Romani, Olson, Semenza, & Grana, 2002). Or, ce type d'analyse peut être influencé par le propre système perceptif de l'expérimentateur (Marczyk & Baqué, 2013).

Afin de pallier ce biais, quelques auteurs ont réalisé des analyses acoustiques des productions des patients en vue de faire émerger des éléments constitutifs de la parole des patients aphasiques de manière objective et d'ainsi mieux caractériser les erreurs de production du langage (Blumstein, Cooper, Goodglass, Statlender & Gottlieb, 1980 ; Buchwald & Miozzo, 2011 ; Frisch & Wright, 2002 ; Marczyk & Baqué, 2013 ; Nespoulous, Baqué, Rosas, Marczyk & Estrada, 2013 ; Ryalls, Provost & Arsenault, 1995). La plupart des études consiste en l'analyse du délai d'établissement du voisement ou Voice Onset Time (VOT). Celui-ci est défini selon Lisker et Abramson (1964) comme l'intervalle de temps entre la détente de l'occlusion de la consonne et le début des vibrations périodiques régulières. Il 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. Ce paramètre 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. Généralement, les études ont montré qu'en raison des difficultés de coordination entre les articulateurs et de réalisation motrice des sons de parole, les patients avec une atteinte phonétique présentaient des difficultés pour marquer la différenciation entre les occlusives voisées et non voisées, pourtant bien sélectionnées au niveau phonologique. En conséquence, les paraphasies phonétiques consistaient en la production d'occlusives dont les VOT s'éloignaient des valeurs prototypiques de leur langue, dans le sens d'un rapprochement entre réalisations voisées et non voisées. En particulier, dans certaines études, les auteurs font état d'un grand nombre de dévoisements d'occlusives voisées chez les patients, en raison de difficultés à maintenir simultanément voisement et occlusion supra-glottique (e.g., Marczyk & Baqué, 2013). Par contre, les études montrent qu'en cas d'atteinte phonologique, les patients présentent des difficultés de sélection des phonèmes dans la catégorie phonologique adéquate au sein du système, entrainant des paraphasies phonologiques, qui consistent autant en des voisements d'occlusives non voisées qu'en des dévoisements d'occlusives voisées, mais dont les moyennes des VOT produits sont toujours proches des valeurs prototypiques observées pour ces deux catégories phonologiques dans la langue du participant (Blumstein et al., 1980 ; Marczyk & Baqué, 2013 ; Nespoulous et al., 2013 ; Ryalls et al., 1995). Cependant, les études portant sur l'analyse acoustique des erreurs de production de patients aphasiques restent peu nombreuses et ont principalement été réalisées en langue anglaise ou espagnole (Blumstein et al., 1980 ; Buchwald & Miozzo, 2011 ; Frisch & Wright, 2002 ; Marczyk & Baqué, 2013 ; Nespoulous et al., 2013). À notre

connaissance, une seule étude s'est centrée sur l'analyse des productions de patients francophones (Ryalls et al., 1995). Cette étude compare les VOT produits par deux groupes de patients ayant des troubles présentés comme soit phonologiques, soit phonétiques. Les résultats n'ont pas montré de différence entre les groupes au niveau des valeurs des VOT. Cependant, les patients étaient classés a priori dans une catégorie de trouble en fonction de l'origine de leur lésion (frontale vs temporale), et il est probable que le lien entre nature du trouble et localisation de la lésion ne soit pas biunivoque (Laganaro, 2015), c'est-à-dire que certains patients présentaient en fait les deux types de troubles à des degrés divers. De plus, cette étude a analysé les performances moyennes de groupes de participants, ce qui peut avoir masqué les différences inter-individuelles, fréquentes dans l'aphasie.

La présente étude s'inscrit dans le cadre d'un projet de recherche dont l'objectif est de caractériser les troubles du langage oral à étiologie cérébrale à l'aide d'outils d'analyse acoustique. Elle consiste en une étude de cas de deux patients aphasiques. D'autres patients sont également en cours d'évaluation à l'heure actuelle et viendront compléter les résultats présentés ici. En vue de caractériser les troubles affectant la production de la parole chez ces patients, nous avons réalisé une analyse acoustique de leurs productions dans une épreuve de répétition d'items bi-syllabiques de type consonne/voyelle, consonne/voyelle, comportant les six consonnes occlusives voisées et non voisées du français /p,t,k,b,d,g/ associées aux trois voyelles cardinales /a,i,u/. En raison de la volonté d'analyser un corpus équilibré, nous avons présenté principalement des non-mots aux patients (ex. /laku/), parmi lesquels se trouvaient certains mots (ex. /papa/). Nous avons choisi de présenter une tâche de répétition car les patients présentaient des troubles importants de la lecture. En ce qui concerne les analyses effectuées, nous nous sommes principalement centrés sur l'analyse du VOT des consonnes occlusives voisées et non voisées. En effet, bien qu'il existe d'autres indices acoustiques liés au contraste de voisement, le VOT est l'indice acoustique principal qui lui est relié et il permet dès lors une comparaison avec la littérature sur le sujet. Nos hypothèses étaient les suivantes : en cas d'atteinte phonétique, nous nous attendions à ce que les patients présentent des difficultés de tenue du voisement des consonnes occlusives voisées. En effet, les occlusives voisées du français présentent un VOT négatif long qui témoigne d'un maintien simultané d'une occlusion supra-laryngée et de voisement au niveau laryngé, nécessitant un contrôle fin des articulateurs concernés étant donné les contraintes aérodynamiques associées. Par conséquent, nous nous attendions à observer un nombre plus important de dévoisements (complets ou partiels), objectivés par des VOT négatifs plus courts en moyenne (Laeufer, 1996). En cas d'atteinte phonologique, la sélection des phonèmes au sein du système serait déficiente. Les paraphasies phonologiques consisteraient dès lors entre autres en des substitutions d'occlusives voisées par des non voisées et des substitutions de non voisées par des voisées, dont les VOT resteraient cependant dans les normes des réalisations observées en langue française. Nous pourrions également noter des substitutions phonologiques qui consisteraient en un changement de lieu d'articulation, en raison d'une sélection erronée d'un phonème dans le système phonologique (cf. Nespoulous et al., 2013).

2 Méthodologie

Notre contribution consiste en l'étude approfondie de deux cas de patients aphasiques, CL et TM. Nous avons réalisé dans un premier temps une analyse de leurs capacités langagières à l'aide de tâches de compréhension et de production du langage fréquemment utilisées en aphasiologie, décrites ci-dessous. Nous avons également évalué leurs capacités auditives ainsi que leurs fonctions exécutives. Enfin, dans le but de déterminer si les erreurs affectant la matière phonique présentes chez les deux patients étaient de nature phonétique ou phonologique, nous avons réalisé une analyse acoustique des productions des patients dans une tâche de répétition de non-mots. Les deux 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 environ 45 minutes à 1 heure. L'ordre était le suivant : Jour 1 : (1) Anamnèse, (2) Description d'images, (3) Dénomination d'images (40 premiers items), (4) Répétition de non-mots (42 premiers items), (5) Exécution de consignes variant en complexité; Jour 2 : (1) Dénomination d'images (40 derniers items), (2) Répétition de non-mots (42 derniers items), (3) Appariement sémantique d'images, (4) Désignation de mots ; Jour 3 : (1) Évaluation des fonctions exécutives, (2) Audiométrie tonale. Les séances d'évaluation ont eu lieu en mars 2015.

2.1 Description des patients

CL est un homme âgé de 65 ans. Il est francophone, droitier et a fait 16 ans d'études. Aujourd'hui retraité, il a été directeur d'une usine de fabrique de clés. En janvier 2013, il est victime d'une hémorragie cérébrale ayant entrainé une lésion cortico-sous-corticale gauche entreprenant une vaste partie du lobe frontal à hauteur des gyri supérieur et moyen et du pôle antérieur du lobe temporal. Il présente alors une aphasie affectant la production et la compréhension du langage ainsi qu'une hémiplégie droite. Depuis, ses capacités de compréhension du langage se sont améliorées mais ses capacités de production restent altérées. Au moment de l'évaluation, CL suit toujours des séances de thérapie langagière à raison de trois heures par semaine. Au niveau visuel, il présente une presbytie corrigée. Le seuil auditif moyen pour ses deux oreilles à 250, 500, 1000, 2000 et 4000 Hz, mesuré en audiométrie tonale liminaire en conduction aérienne est de 27.19 dB. L'évaluation neuropsychologique de ses capacités exécutives, mesurées à l'aide de tests ne faisant pas intervenir le langage (Quinette et Lambert, 2013), indique la présence de difficultés de mémoire à court terme et de travail, d'inhibition et de mise à jour. L'évaluation du langage oral du patient a montré que CL ne présentait pas de trouble de la compréhension du langage en désignation d'images (*Examen Long du Langage, UCL/ULg*), en exécution de consignes variant en complexité (*Token Test*, De Renzi, 1962) ainsi qu'au niveau sémantique, mesuré par une tâche d'appariement sémantique d'images (*Pyramids and Palm Trees Test*, Howard & Patterson, 1992). Par contre, les capacités de production du langage sont altérées chez le patient. En description d'image (*The cookie theft picture,* Goodglass et al., 2000), le patient présente un langage hésitant, avec de nombreuses pauses à des endroits inappropriés dans la phrase (e.g., en milieu de phrase). En dénomination d'images (*Lexis*, de Partz et al., 2001), CL présente un score de 74% de réponses correctes. Il montre un effet de longueur mais pas d'effet de fréquence. On note la présence de quelques paraphasies sémantiques (ex. jupe → « robe »), visuo-sémantiques (ex. train → « locomotive ») et des non-réponses, mais principalement de paraphasies atteignant la forme phonique des mots (ex. pantalon → [panatɔ̃]), allant parfois dans le sens d'un dévoisement (ex. lampadaire → [lɑ̃patɛʁ]). En répétition de syllabes variant en complexité articulatoire (*Examen Long du Langage*), CL répète correctement 73% des syllabes et présente un effet de complexité articulatoire. En répétition de mots variant en complexité articulatoire et en longueur (*Examen Long du Langage*), le patient répète correctement 83% des mots et présente un effet de longueur. Ses erreurs consistent en des paraphasies atteignant la forme phonique des items. L'ensemble de ces résultats nous a amenés à faire l'hypothèse de la présence de difficultés de type phonologiques ou phonétiques chez CL.

TM est un homme âgé de 62 ans. Il est francophone, droitier et a fait 16 ans d'études. Aujourd'hui retraité, il a été professeur d'éducation physique. En avril 2006, il est victime d'une hémorragie cérébrale ayant entrainé une lésion cortico-sous-corticale gauche au niveau fronto-pariétal. Il présente alors une aphasie affectant la production et la compréhension du langage ainsi qu'une hémiplégie droite et une hémiparésie faciale gauche. Il a bénéficié de deux ans de thérapie langagière jusqu'en 2008. Ses capacités de compréhension du langage se sont améliorées mais ses capacités de production restent altérées. Depuis son accident vasculaire en 2006, il présente de l'épilepsie, aujourd'hui

contrôlée. En 2009, il a également été victime d'un deuxième accident vasculaire cérébral, qui n'a pas entraîné de complication supplémentaire. Au niveau visuel, on note une presbytie corrigée et son seuil auditif moyen pour les deux oreilles et de 18.75 dB. Au niveau de ses capacités exécutives, le patient présente des difficultés de mémoire à court terme et de travail, d'inhibition et de mise à jour. En ce qui concerne ses capacités en langage oral, nos analyses indiquent que TM ne présente pas de difficulté de compréhension du langage oral. Par contre, ses capacités de production du langage sont altérées. En description d'image, le patient présente de nombreux épisodes de manque du mot marqués par des pauses à des endroits inappropriés. Il commet des erreurs au niveau de la restitution de la forme phonique des mots (ex. éclabousse → [ekabus]). En dénomination d'images, TM présente 66% de réponses correctes. Il montre un effet de longueur et de fréquence. Il commet principalement des paraphasies de type sémantique (assiette → « bol ») et au niveau phonique allant dans le sens de substitutions ou de suppressions de phonèmes dans des groupes consonantiques (édredon → [erədɔ̃], panier → p̲n̲ja]). On note également la présence de quelques paraphasies visuo-sémantiques (renne → corne) ainsi que des non-réponses. En répétition de syllabes variant en complexité articulatoire (*Examen Long du Langage*), TM répète correctement 93% des syllabes. En répétition de mots variant en complexité articulatoire et en longueur (*Examen Long du Langage*), le patient répète correctement 77% des mots et présente des effets de longueur et de complexité articulatoire. Ses erreurs consistent en des paraphasies atteignant la forme phonique des items. Les résultats de TM indiquent la présence de difficultés lexico-sémantiques ainsi que phonologiques ou phonétiques.

2.2 Tâche de répétition de non-mots

Notre intérêt s'articulant autour du trait de voisement en langue française, nous avons choisi 84 non-mots CVCV, comprenant les occlusives voisées et non voisées du français /p,t,k,b,d,g/ ainsi que les voyelles cardinales du français /a,i,u/. À savoir 18 items C1V1$C_{[l]}V_{[a]}$ où C1=/p,t,k,b,d,g/ et V1=/a,i,u/ (p.ex. /pula/); 18 items $C_{[l]}V_{[a]}$C2V2 où C2=/p,t,k,b,d,g/ et V2=/a,i,u/ (p.ex. /lapu/); 36 items C1$V_{[a]}$C2$V_{[a]}$ où C1 et C2 = /p,t,k,b,d,g/ (p.ex. /gada/) et 12 items C1V1C2V2 où C1=C2=/p,t,k,b,d,g/ et V1=V2=/i,u/ (p.ex. /kiki/). Les items ont été préalablement enregistrés par une locutrice francophone en chambre sourde 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, accompagnés de la phrase « ce sont des » afin de limiter l'impact éventuel de difficultés d'initiation de la parole chez les patients. La tâche consistait en une répétition des non-mots insérés au sein de la phrase porteuse (ex. « Ce sont des /kidi/ »). Les productions des patients ont été enregistrées (enregistreur audio portable Zoom H5 avec couple stéréo en X/Y) en vue d'une analyse ultérieure.

3 Résultats

Nous avons procédé à une analyse acoustique des productions du patient à l'aide du logiciel Praat (Boersma & Weenink, 2009). Nous nous sommes centrés sur la mesure des VOT, qui ont été mesurés manuellement (en ms) sur l'oscillogramme comme l'intervalle temporel entre le début du burst et le début du voisement (pour la voyelle suivante, ou, le cas échéant, au cours de l'occlusion). Afin d'objectiver la présence d'un voisement pour les occlusives voisées, nous nous sommes centrés sur la présence de pulsations périodiques avant l'apparition de l'explosion de la consonne, marquées par la présence de barres de voisement régulières sur le spectrogramme. En cas d'absence de périodicité avant l'explosion de l'occlusive, nous avons considéré la consonne comme non voisée. En cas de présence de périodicité, nous avons pris la première de ces pulsations périodiques comme référence pour la mesure du VOT (négatif), même si l'ensemble de l'occlusive n'était pas voisée (voir ci-dessous). Nous avons tout d'abord calculé les valeurs moyennes des VOT pour les deux patients, en fonction du type d'occlusive attendue (sollicitée par le corpus) et de la présence effective ou non de

périodicité dans leurs productions (cfr. Table 1). Nous avons également calculé le pourcentage d'occlusives présentant de la périodicité alors qu'elle était attendue (productions correctes) ou non (erreurs). Notons que pour cette analyse, nous nous sommes centrés sur les consonnes dont le lieu d'articulation restait identique au lieu attendu. Nous avons donc exclu de ces analyses les consonnes qui étaient substituées par des consonnes dont le lieu d'articulation était différent (ex. p → [t]); il s'ensuit que la somme des pourcentages dans la table 1 n'équivaut pas toujours à 100. Comme nous pouvons le constater, CL présente des valeurs moyennes de VOT dans les normes pour la langue française pour les consonnes occlusives voisées (normes : -100 ms, Laeufer, 1996) et non voisées (normes : +30 ms, Laeufer, 1996). Néanmoins, dans 41% des cas où CL doit produire une occlusive voisée, il réalise une occlusive non voisée. Chez TM, les valeurs moyennes des VOT des occlusives non voisées sont également dans les normes. Par contre, les valeurs moyennes des VOT pour les occlusives voisées (-69 ms) sont supérieures à celles généralement rencontrées dans la langue française. De plus, ce voisement ne survient que dans 55% des cas (vs. 21% de dévoisements complets). Il voise également 30% des non voisées.

	CL		TM	
Type occlusive attendue	Périodicité	VOT moyen (ms)	Périodicité	VOT moyen (ms)
Voisée	Oui (59%)	-98.00	Oui (55%)	-69.00
	Non (41%)	25.99	Non (21%)	32.22
Non voisée	Non (97%)	30.25	Non (52%)	30.94
	Oui (3%)	-101.20	Oui (30%)	-70.10

TABLE 1 : Valeurs moyennes des VOT pour les patients CL et TM et pourcentages de présence ou d'absence de périodicité en fonction du type d'occlusive attendue.

Nous nous sommes ensuite intéressés à la nature des erreurs produites par les participants, à savoir, pour chaque patient, la proportion des trois types d'erreurs (dévoisements de consonnes occlusives voisées; voisements d'occlusives sourdes; substitutions de phonèmes consistant en un changement de lieu d'articulation) sur le total des erreurs commises, en fonction de la position de la syllabe concernée dans le non-mot. Ces résultats sont résumés dans la Figure 1.

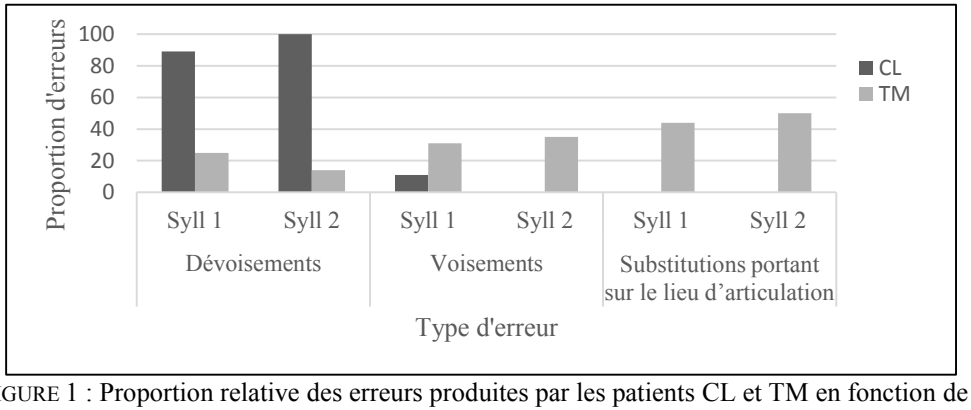


FIGURE 1 : Proportion relative des erreurs produites par les patients CL et TM en fonction de la position de la syllabe dans le non-mot: dévoisements d'occlusives voisées, voisements d'occlusives non voisées, substitutions portant sur le lieu d'articulation.

Les résultats indiquent que la grande majorité des erreurs commises par CL consiste à dévoiser les occlusives sonores, que ce soit dans la première syllabe ou dans la deuxième. Il commet par contre très peu de voisements de consonnes sourdes et aucune substitution consistant en un changement de lieu articulatoire. Notons que la tendance aux dévoisements dans la deuxième syllabe survient

principalement lorsque la consonne de la première syllabe est non voisée (ex. pada → [pata]). TM ne montre pas de tendance préférentielle à dévoiser les occlusives sonores ou à voiser les occlusives sourdes. En effet, le pourcentage d'erreurs de ces types est relativement similaire. Par contre, nous notons une proportion importante d'erreurs de substitutions qui consistent en un changement de lieu d'articulation des occlusives concernées.

En plus des mesures acoustiques des VOT, notre analyse acoustique a également mis en évidence des phénomènes intéressants, visibles sur les spectrogrammes mais peu, voire non détectables sur base d'une évaluation purement perceptive des enregistrements. Des illustrations sont présentées à la figure 2. Ainsi, chez CL, lors de la production de certaines consonnes voisées, nous avons noté la présence d'un arrêt du voisement dans la deuxième partie de l'occlusion, quelques (dizaines de) millisecondes avant l'explosion de la consonne (voir [daba] pour ce patient dans la Figure 2). Ces résultats semblent indiquer la présence de difficultés à maintenir voisement et occlusion simultanément chez ce patient. De plus, chez les deux patients, nous notons la présence d'un allongement de la voyelle précédant une consonne voisée qui présentait également une explosion plus faible. Une interprétation potentielle de ce phénomène est qu'il est le résultat d'un mécanisme mis en place par les patients dans le but de compenser les difficultés de voisement des occlusives voisées (voir [bubu] pour CL et [bibi] pour TM).

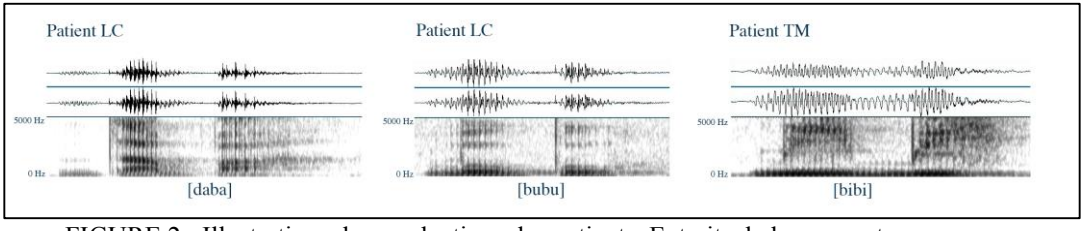


FIGURE 2 : Illustrations des productions des patients. Extraits de leurs spectrogrammes.

4 Discussion

En aphasiologie, il est indispensable de caractériser les erreurs de production du langage afin d'émettre des hypothèses précises quant au locus d'altération responsable des troubles du patient. Cela permet de programmer de manière efficace le travail de rééducation qu'il faudra mener avec lui. Néanmoins, en ce qui concerne les erreurs de production des phonèmes, il est parfois malaisé de déterminer dans quelle mesure le patient présente une atteinte de la sélection des phonèmes au sein du système phonologique (atteinte phonologique) ou plutôt un déficit de la programmation motrice ou de la réalisation articulatoire des sons de parole (atteinte phonétique). En effet, non seulement les erreurs sont parfois très similaires en surface, mais leur appréciation est généralement basée sur une analyse perceptive réalisée par l'expérimentateur, qui peut être biaisée par son propre système perceptif. Quelques auteurs ont certes proposé une analyse acoustique des erreurs du patient dans le but de mettre en évidence de manière objective des éléments de la production de parole chez les aphasiques (Blumstein et al., 1980 ; Buchwald & Miozzo, 2011 ; Frisch & Wright, 2002 ; Marczyk & Baqué, 2013 ; Nespoulous et al., 2013 ; Ryalls et al., 1995). Cependant, les études utilisant ce type de méthodologie restent rares et, dans le domaine francophone, quasi inexistantes.

Notre étude a porté sur l'analyse des erreurs de réalisation des occlusives voisées et non voisées de deux patients aphasiques francophones, CL et TM, engagés dans une tâche de répétition de non-mots. En effet, dans des tâches classiques de description et de dénomination d'images, nous avions objectivé la présence d'erreurs affectant la forme phonique des mots chez les deux patients. Les résultats ont

montré que CL présentait des valeurs de VOT dans les normes de la langue française pour les occlusives voisées et non voisées. Par contre, l'analyse des erreurs produites par le patient a montré une tendance marquée à dévoiser les occlusives voisées (41% d'entre elles sont complètement dévoisées). Bien que ces difficultés pourraient être interprétées comme résultant d'une atteinte phonologique (substitution), le fait que nous ne notions chez CL qu'un très faible pourcentage de voisements de consonnes sourdes et aucune erreur consistant en un changement de lieu articulatoire irait, selon la littérature, dans le sens de la présence de troubles phonétiques chez ce patient (Blumstein et al., 1980 ; Marczyk & Baqué, 2013 ; Nespoulous et al., 2013 ; Ryalls et al., 1995). En outre, nous avons noté la présence d'arrêts du voisement avant l'apparition de l'explosion de la consonne lors de la production d'occlusives voisées chez ce patient (dévoisements partiels). Bien que ces phénomènes soient également présents chez l'adulte sain (Ivent, Adda-Decker, & Fougeron, 2015), leur fréquence importante chez ce patient et le contexte de difficultés articulatoires nous conduit à les interpréter comme étant un signe de difficultés à coordonner adéquatement les articulateurs laryngés et supra-laryngés afin de maintenir simultanément voisement et occlusion supra-glottique. Enfin, nous avons également remarqué que certaines erreurs de dévoisement d'occlusives voisées situées dans la deuxième syllabe pourraient être dues à des persévérations du dévoisement de la première occlusive non voisée. Ce phénomène pourrait être lié aux difficultés exécutives du patient, mises en évidence dans l'évaluation de ses fonctions exécutives, en l'occurrence des difficultés d'inhibition des items/phonèmes/syllabes présentés précédemment. Chez TM, nous avons noté pour les occlusives voisées un VOT négatif moins long, en moyenne, que celui généralement observé chez des locuteurs francophones (alors que son débit de parole n'est pas plus rapide); cette courte durée d'occlusion voisée est concomitante à un allongement (compensatoire?) de la voyelle précédant la consonne cible. Selon la littérature, ce phénomène pourrait indiquer la présence de difficultés de la tenue du voisement en raison d'une atteinte phonétique (Blumstein et al., 1980 ; Marczyk & Baqué, 2013 ; Nespoulous et al., 2013 ; Ryalls et al., 1995). Enfin, l'analyse des types d'erreurs réalisées par le patient montre que TM produit un pourcentage plus élevé de substitutions (en termes de lieu d'articulation) que d'erreurs de voisements des occlusives. Ces erreurs de substitution peuvent être dues à des difficultés de sélection des phonèmes adéquats au sein du système phonologique, indiquant la présence de troubles phonologiques chez TM. Cependant, il est également possible que les difficultés exécutives mises en évidence chez le patient soient en partie responsables de ces erreurs.

Les résultats de cette étude montrent l'intérêt d'adjoindre des analyses acoustiques des productions des patients aux évaluations perceptives le plus souvent réalisées dans le but de caractériser leurs erreurs. En effet, certains phénomènes ne sont pas aisément perceptibles sans ces analyses, tel que cela avait été démontré dans des études précédentes, en langue anglaise ou espagnole (Blumstein et al., 1980 ; Buchwald & Miozzo, 2011 ; Frisch & Wright, 2002 ; Marczyk & Baqué, 2013 ; Nespoulous et al., 2013 ; Ryalls et al., 1995). En outre, nos analyses montrent que les troubles phonétiques présents chez les deux patients ne se manifestent pas de la même manière selon le patient, et que les productions d'un même patient pour un même phonème sont également très variables d'une production à l'autre. Ces observations soulignent par conséquent l'importance de procéder à des études de cas unique plutôt qu'à des comparaisons de groupes de patients qui risquent de masquer ces différences intra- et inter-individuelles (p.ex., Ryalls et al., 1995). De plus, il est important de prendre en compte la présence de stratégies palliatives mises en place par les patients dans le but de compenser leurs difficultés phonétiques. En effet, nous avons observé chez CL et TM la présence d'un allongement de la voyelle précédant une occlusive voisée à l'attaque de la syllabe suivante, allongement qui pourrait avoir pour effet de compenser partiellement le manque de voisement de celle-ci. La covariation entre voisement des occlusives et durée des voyelles adjacentes est un phénomène bien connu dans les langues du monde (Kluender, Diehl, & Wright, 1988). La présence de stratégies palliatives en vue de compenser les difficultés de planification ou de réalisation

articulatoires a déjà été signalée dans la littérature en langue espagnole (Marczyk & Baqué, 2013 ; Nespoulous et al., 2013). Ceci se marquait par une augmentation de la tension dans les consonnes occlusives, objectivée par une augmentation des barres d'explosion et de la durée de l'explosion pour ces consonnes. Selon les auteurs, étant donné que les patients utilisent des mécanismes palliatifs en vue de compenser leurs difficultés de gestion du voisement, cela signifierait que la distinction entre les occlusives voisées et non-voisées serait effective au niveau phonologique, mais que la réalisation phonétique serait déficiente. Ce type d'hypothèse pourrait également s'appliquer à nos observations. En effet, les mécanismes de compensation des difficultés de voisement de l'occlusive pourraient indiquer que l'occlusive voisée a correctement été sélectionnée au niveau phonologique mais que la réalisation motrice du trait de voisement serait déficitaire pour les patients. Enfin, nos analyses suggèrent également que d'autres mécanismes cognitifs interfèrent avec les mécanismes de production langagière des patients, tels que les difficultés exécutives, qui semblent être responsables de certaines erreurs dans la réalisation des phonèmes produits par CL et TM. Cela avait déjà été observé dans la littérature, au niveau des erreurs de type lexico-sémantique (e.g., Martin & Allen, 2008).

Par ailleurs, d'un point de vue méthodologique, cet article souligne la pertinence de la mise en place, au-delà du VOT, d'un plus grand nombre de critères basés sur des paramètres acoustiques précis en vue d'objectiver les phénomènes observés. En effet, les productions des patients étaient très variables et déviaient de la norme attendue sur de multiples dimensions. Dans le cas des mécanismes compensatoires précités, par exemple, nos observations sont à ce stade fondées sur un examen approfondi des signaux au cas par cas, sans mesure *ad hoc*. La même critique peut également s'appliquer à la mesure du voisement utilisée, basée sur la présence d'une périodicité sur le signal pendant la phase d'occlusion, même si cette périodicité est interrompue avant l'explosion. Dans la suite de nos travaux, nous souhaiterions systématiser nos observations par l'utilisation d'indices à base acoustique (tels, par exemple, le taux de passages à zéro ou le Noise-to-Harmonic Ratio) dans le but d'objectiver plus précisément la présence, fût-elle stable ou intermittente, de périodicité tout au long de l'occlusion. Ceci permettrait d'appréhender le phénomène de manière moins dichotomique et de mettre en évidence la présence de signaux faibles et difficilement perceptibles mais néanmoins susceptibles d'attester une intention de voisement de la part du patient. Ceci constituerait un indice supplémentaire pour déterminer si, sur le plan phonologique, la consonne voisée a bien été sélectionnée, quand bien même le résultat conduirait l'observateur à conclure à l'absence de voisement. En outre, nous pourrions également analyser la présence d'autres indices acoustiques tels que la durée des voyelles adjacentes, connus pour faire partie de la réalisation du contraste de voisement.

En conclusion, notre étude souligne l'intérêt de l'utilisation d'analyses acoustiques dans le but d'aider au diagnostic différentiel entre les erreurs phonétiques et phonologiques dans l'aphasie, élément essentiel pour le choix de la méthode de rééducation langagière du patient. En outre, les résultats montrent une grande diversité des manifestations des troubles langagiers des patients, ce qui démontre l'intérêt de procéder à des analyses précises à l'aide de critères multiples, susceptibles d'étayer des raisonnements métrologiques et contribuant, dès lors, à l'objectivation des phénomènes observés.

Remerciements

Cette recherche a été subventionnée par l'Action de Recherche Concertée ParolPathos (AUWB-2012-12/17-UMONS- N°1).

Nous remercions Florence Piertot, Diane Lecat, Valérie Chavet et Jérémy Pouliart pour l'aide apportée dans la récolte des données.

Références

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.

BUCHWALD A & MIOZZO M. (2011). Finding levels of abstraction in speech production: evidence from sound-production impairment. *Psychological Sciences*, **22**, 1113-1119.

CHO, T. & LADEFOGED, P. (1999). Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics*, **27**, 207-229.

FRISCH, S.A. & WRIGHT, R. (2002). The phonetics of phonological speech errors: An acoustic analysis of slips of the tongue. *Journal of Phonetics*, **30**, 139–162.

KLUENDER, K., DIEHL, R., & WRIGHT, B. (1988). Vowel-length differences before voiced and voiceless consonants: An auditory explanation. *Journal of Phonetics*, **16**, 153-169.

IVENT, F., ADDA-DECKER, M., & FOUGERON, C. (2015). Voicing variations in French obstruents: Distribution and acoustic quantification. [Abstract]. *18th International Congress of Phonetic Sciences (ICPhS'15)*, 5.

LAEUFER, C. (1996). The acquisition of a complex phonological contrast: voice timing patterns of English initial stops by native French speakers. *Phonetica*, **53**, 86–110.

LAGANARO, M. (2015). Paraphasies phonémiques et/ou phonétiques ? Des raisons et des difficultés de cette distinction. *Revue de neuropsychologie*, **7**, 27-32.

LEVELT, W. J., ROELOFS, A., & MEYER, A. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, **22**(1), 1-75.

LISKER, L., & ABRAMSON, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, **20**, 384-422.

MARCZYK, A., & BAQUÉ, L. (2013). De l'origine des erreurs de substitution consonantique chez les patients aphasiques hispanophones : une étude acoustique. *Recherches en Parole : La voix et la parole perturbées, Travaux en Phonétique Clinique,* **1**(1), 157-170.

MARTIN, R.C., & ALLEN, C.M. (2008). A Disorder of Executive Function and Its Role in Language Processing. *Seminars in Speech and Language*, **29**(3), 201-210.

NESPOULOUS, J.-L., BAQUÉ, 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**, 117-125.

RYALLS, J., PROVOST, H., & ARSENAULT, N. (1995). Voice onset time production in French-speaking aphasics. *Journal of Communication Disorders*, **28**, 205-215.

ROMANI, C., OLSON, A., SEMENZA, C., & GRANÀ, A. (2002). Patterns of phonological errors as a function of a phonological versus an articulatory locus of impairment. *Cortex*, **38**(4), 541-567.

Dynamique phonétique et contrôle moteur dans la maladie de Parkinson: analyse du contrôle de la production des glides

Virginie Roland[1,2], Véronique Delvaux[1,3], Kathy Huet[1], Myriam Piccaluga[1], Marie-Claire Haelewyck[2], 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) Service d'Orthopédagogie Clinique, Université de Mons, Belgique
(3) Fonds National de la Recherche Scientifique, Belgique

Virginie.roland@umons.ac.be

RESUME

Nous nous interrogeons quant à la possibilité d'identifier les difficultés de contrôle du mouvement chez les personnes atteintes de la maladie de Parkinson (MP) à partir de l'étude de leurs comportements dans la production de sons de parole nécessitant des mouvements continus des articulateurs supralaryngés (logatomes VCV, où C est un glide). Notre hypothèse est que les parkinsoniens présentent des modifications dans leur dynamique de mouvement par rapport à des personnes sans pathologie lors de la production. A cette fin, sont étudiés des sons de parole recueillis hors contexte communicationnel auprès de neuf personnes porteuses de la MP et de dix sujets sains. Les analyses révèlent des différences entre les deux groupes, notamment en ce qui concerne l'espace articulatoire, l'amplitude des mouvements et leur localisation dans le plan F_1-F_2. On note par ailleurs qu'un point-cible est préservé lors de l'émission de logatomes : le centre du glide.

ABSTRACT

Speech dynamics and motion control in people with Parkinson's disease: analysis of glides' production

We wonder about the possibility of identifying motion control disorders in people with Parkinson's disease (PD) from the study of their behavior in the production of speech sounds requiring continuous movements of the supralaryngeal articulators (logatomes VCV, where C is a glide). Our hypothesis is that Parkinsonians differ in their movement dynamic relative to people without pathology during speech production. To this end, speech sounds were collected from nine people carrying the PD and ten healthy subjects. The analyzes reveal differences between the two groups, in particular as regards the articulatory space, range of motion and location in the F_1-F_2 plane. We also note that a target point is preserved during production of logatomes: the center point of the glide.

MOTS-CLES : dynamique de parole, maladie de Parkinson, glide, analyse acoustique
KEYWORDS: speech dynamics, Parkinson's disease, glide, acoustic study

1 Introduction

La maladie de Parkinson (désormais MP) est une pathologie neurodégénérative se traduisant par une perte progressive de neurones de la voie nigro-striée, entrainant un déficit fonctionnel du striatum et, de ce fait, une perte progressive en dopamine striatale. Cette dégénérescence induit d'importantes difficultés dans le contrôle des mouvements, pouvant prendre diverses formes, d'où la notion de « triade parkinsonienne », englobant tremblements de repos (symptôme initial dans 70% des cas), akinésie et rigidité (avec association de troubles posturaux et de la marche). D'autres symptômes peuvent également être associés à la maladie, notamment des signes axiaux. Bien que d'apparition souvent tardive, ces derniers peuvent s'avérer in fine envahissants. La dysarthrie en est part intégrante. L'étiologie neuro-dégénérative de la maladie conduit à une grande variété de troubles de la parole généralement regroupés sous l'étiquette de « dysarthrie hypokinétique ». Celle-ci, telle que définie par Darley *et al.* (1975) se manifeste dans tous les aspects de la production de la parole, avec des répercussions sur les processus respiratoires, phonatoires et articulatoires, tant au niveau segmental que suprasegmental. Des études classiques à base perceptuelle (Darley *et al.*, 1975 ; Logemann *et al.*, 1978) ainsi que des études acoustiques au cours des vingt dernières années (Gamboa *et al.*, 1997 ; Cheang & Pell, 2007) ont montré à maintes reprises que les locuteurs atteints de la MP présentent des troubles de la qualité de la voix (voix rauque et soufflée, etc.), une variation de l'émission limitée en intensité et en fréquence fondamentale (monotonie de hauteur et d'intensité), une variabilité du débit de parole non contrôlée, comprenant notamment des pauses plus longues, inappropriées, et des répétitions de mots et/ou de syllabes. Sur le plan articulatoire, la production imprécise de consonne est l'un des éléments les plus souvent signalés (Ackermann & Ziegler, 1991 ; Wong *et al.*, 2011). Ce sont notamment les consonnes occlusives, fricatives et affriquées qui présentent le plus de distorsions, probablement en raison de la réduction de l'amplitude et de la force du mouvement articulatoire. Ackermann *et coll.* (Ackermann & Ziegler, 1991 ; Ackermann *et al.*, 1995) ont d'ailleurs émis l'hypothèse que les personnes atteintes de la MP réduisent l'amplitude de leurs mouvements articulatoires afin de préserver le tempo de parole, ce qui conduit à un phénomène d'hypoarticulation. Cependant, des études physiologiques ne corroborent que partiellement ces constats, en ce qui concerne l'amplitude et la vitesse de mouvement de la mâchoire, de la langue et des lèvres (ainsi que de l'activité musculaire associée) dans la production de la parole parkinsonienne (Walsh & Smith, 2012).

Beaucoup des études interrogeant les répercussions vocales de la maladie de Parkinson procèdent par analyse de segments de parole. Il s'agit tantôt de voyelles (Bang, Min, Sohn, & Cho, 2013 ; Hertrich, Lutzenberger, Spieker & Ackermann, 1997), tantôt de consonnes (Cf. supra). Dans diverses recherches, les corpus sont constitués de sons de parole connectés, ce qui permet l'étude de la coarticulation (Tjaden, 2000 ; Tjaden & Sussman, 2006). Plus récemment, une attention soutenue a été réservée aux tendances prosodiques de la parole signifiante (Duez, Jankowski, Purson, & Viallet, 2012 ; Ghio, Robert, Grigoli, Mas, Delooze, Mercier, & Viallet, 2014). Si maintes études se sont ainsi basées, à l'origine, sur des productions assez statiques, force est de reconnaître que la dynamique de production de la parole a plus récemment fait l'objet d'une attention renforcée.

Nonobstant, c'est la plupart du temps avec une focalisation soit sur la variabilité intrinsèque du signal laryngé soit sur des variations inter-segmentales qu'est approchée la dynamique du signal acoustique. Comme le remarquent Goberman et Coelho (2002), peu de recherches se sont concentrées sur la dynamique intra-segmentale du timbre imputable au contrôle des résonateurs supra-glottiques, dont l'étude est pourtant parfaitement adaptée à la MP. Moins de travaux encore se sont centrés sur les caractéristiques des segments susceptibles de porter *en eux-mêmes* la trace d'une attente normative

de variabilité du timbre ; or, ceux-ci requièrent, de la part du locuteur, la réalisation de mouvements articulatoires intra-segmentaux très rapides et étroitement contrôlés.

En français, les phonèmes /w/, /ɥ/ et /j/ sont porteurs d'attentes de ce type[1]. Ces sons de parole se caractérisent en effet par l'évolution continuelle de leur timbre au cours de leur production : de qualité acoustique proche, en son début et à son terme, de celle des sons du contexte immédiat, le glide approche, en sa partie médiane, la qualité d'une voyelle du système (/u/ pour /w/, /y/ pour /ɥ/ et /i/ pour /j/).

Comme l'ont précédemment suggéré Harmegnies *et coll.*, les caractéristiques acoustiques des réalisations de ces phonèmes peuvent donc faire figure de matériau de choix pour l'étude du contrôle du mouvement dans la dysarthrie parkinsonienne (Couvreur *et al.*, 1999). C'est cette idée que nous tenterons de mettre à l'épreuve dans cette communication à caractère exploratoire et méthodologique, en testant l'hypothèse que les personnes atteintes de la MP présentent des modifications dans leur dynamique de production des glides par rapport aux sujets exempts de pathologie.

2 Méthodologie

2.1 Locuteurs

Deux groupes de sujets ont été constitués. Le premier se compose de 9 personnes atteintes de la MP. Ces sujets (6 hommes et 3 femmes), tous locuteurs natifs du français de Belgique (Brabant wallon), sont âgés de 52 à 77 ans (âge moyen de 65 ans). Ils présentent une durée moyenne de maladie de 9 ans ; ils occupent globalement une position médiane dans la classification de Hoehn et Yahr (1967) et vivent de manière autonome. Tous sont sous traitement médicamenteux et tous les enregistrements ont été effectués sous médication. Une personne a eu recours à la chirurgie de stimulation cérébrale profonde 5 ans avant le recueil de données ; seule cette personne suit également un traitement logopédique. Les sujets se sont vu appliquer le Voice handicap Index (1997). Leurs résultats vont d'un handicap global léger (HP – GF – YMS) à sévère (PJ – PR – LC – BD). Deux personnes attestent quant à elles d'un handicap vocal modéré (JC – JG). Trois sujets (HP, GF et YMS) ne présentent pas de handicap vocal en ce qui concerne le domaine physique. Aucun sujet ne fait état de quelque plainte que ce soit au niveau articulatoire. Le deuxième groupe est constitué d'un échantillon occasionnel équilibré en genre de 10 sujets contrôles, exempts de toute pathologie.

2.2 Corpus

Les sons de parole recueillis auprès des sujets sont le résultat de productions sollicitées dans un contexte non-communicationnel, à partir de versions écrites des productions à réaliser, apparaissant en transcription orthographique française sur un écran d'ordinateur portable placé face aux participants ; chaque injonction visuelle était accompagnée d'un exemple sonore préenregistré afin que chacun soit soumis à la même version. Chaque locuteur a été prié de réaliser, d'abord, en voyelle tenue, une production stable en isolation de chacune des 3 voyelles orales périphériques de l'espace vocalique (/a/, /i/, /u/) et ensuite des logatomes de structure V_1CV_2 (où C est l'un des glides: /aja/,

[1] Nous centrant sur les caractéristiques acoustiques de leurs réalisations, nous éviterons, à leur propos, les dénominations telles que « semi-voyelle » ou « semi-consonne », afin de contourner tout risque de controverse et retiendrons l'appellation anglophone « glide », qui est bien en ligne avec leurs spécificités phonétiques.

/aju/, /uju/, /awi/ et /awa/). Le recours aux phrases porteuses a été écarté en vue d'éviter le biais d'hypoarticulation susceptible de survenir dans la MP lorsque le sujet est amené à réaliser des productions longues (Cf. Sauvageau, Roy, Cantin, Prud'Homme, Langlois, & Macoir, 2015). Par ailleurs, le phonème /ɥ/ a été exclu, vu sa propension à être confondu avec /w/ dans le régiolecte investigué.

2.3 Analyses acoustiques

Les valeurs des premier et deuxième formants ont été évaluées pour toutes les productions recueillies dans chacun des deux groupes. Pour les voyelles tenues, les mesures ont été effectuées en milieu de tenue. Pour les logatomes, elles ont été pratiquées à l'entame de la réalisation (début de V_1), au point d'inflexion des trajectoires formantiques (au cours de C) et en fin de production (fin de V_2), soit, respectivement, aux points D, I et A sur le partie droite de la figure 1. Les valeurs recueillies, dans un premier temps, au moyen du logiciel de tracking formantique de PRAAT ont été vérifiées et le cas échéant corrigées par recours à l'examen spectrographique. Dans un deuxième temps, toutes les mesures ainsi obtenues ont été vérifiées par un phonéticien chevronné non impliqué dans la campagne de mesure initiale et chaque désaccord a fait l'objet d'une discussion sur la base d'un retour en commun sur les enregistrements analysés.

2.4 Production de données topologiques pour l'étude du contrôle moteur

En perspective du traitement, qui a pour objectif principal d'étudier le contrôle moteur des locuteurs, nous avons opté pour un regard topologique. Chaque timbre de production recueillie a été considéré comme un objet dans le plan F_1/F_2 (Cf. Fig. 1). S'agissant des voyelles tenues, chacune correspond à un point. Pour les logatomes, chacun correspond à un trajet, caractérisable par trois points remarquables : le lieu correspondant à l'entame de la réalisation (début de V_1 : « départ »), celui correspondant à son terme (fin de V_2: « arrivée ») et enfin celui correspondant à l'inflexion des traces formantiques sur le tracé spectrographique (C : « inflexion »).

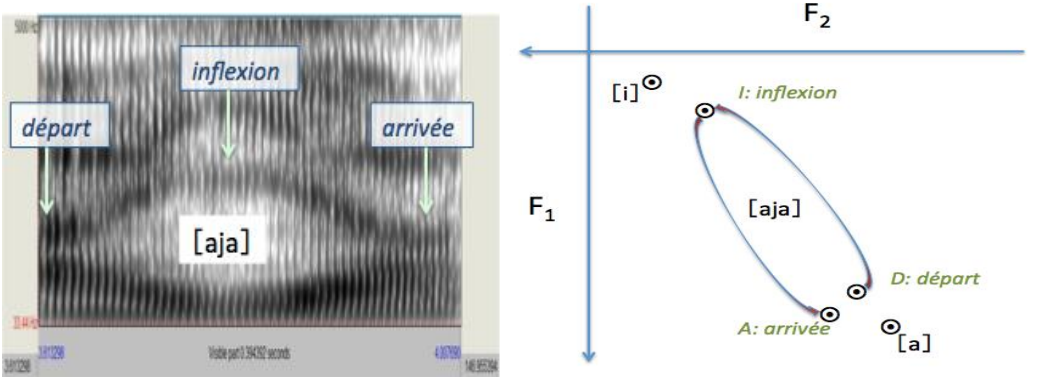


FIGURE 1 : Représentation schématique de la réalisation d'un logatome de type /aja/ : visées spectrographique, à gauche et topologique, à droite (les tracés curvilinéaires représentent la variation constante du timbre durant la production de /aja/, de D à A en passant par I ; les points isolés (« [a] » et « [i] ») représentent les timbres des voyelles tenues produites en isolation

Les données disponibles ont permis le calcul de diverses distances euclidiennes dans le plan F_1/F_2. Nous nous centrerons ici sur deux de leurs types : 1° des distances informant sur l'ampleur de la modification du timbre au sein du logatome (distance « aller » de D à I et distance « retour » de I à A) et, 2°, des distances informant sur la localisation du logatome dans le plan F_1/F_2 par rapport aux voyelles tenues (p.ex. distances Départ-[a], Arrivée-[a] et Inflexion-[i]). Dans le premier cas, les observations sont susceptibles de fournir des informations sur l'amplitude des mouvements articulatoires développés pour la production du glide. Dans le second, les distances permettent d'apprécier l'éventuel déplacement des points D, I et A par rapport aux voyelles tenues correspondantes, et ainsi, de spéculer sur d'éventuelles stratégies articulatoires de production des glides qui différeraient de celles caractérisant la production de monophtongues similaires.

3 Résultats

Dans cette section, nous procédons systématiquement à la comparaison des observations opérées dans le groupe MP à celles provenant du groupe témoin. L'émergence de différences sera ici interprétée comme indiciaire d'un comportement particulier des sujets MP par rapport à la normale.

3.1 Le champ articulatoire

L'analyse des voyelles tenues permet d'apprécier l'étendue du champ articulatoire des sujets en condition de production de monophtongues. Un simple examen descriptif (Cf. fig 2) suggère une variation sensible de la cohésion inter-sujet, l'espace articulatoire utilisé apparaissant beaucoup plus variable chez les parkinsoniens que chez les sujets sains.

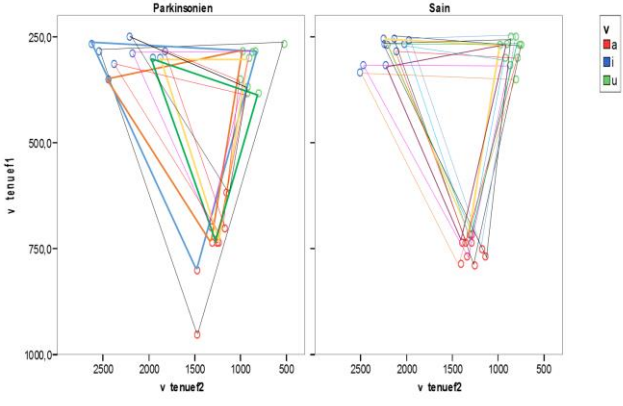


FIGURE 2 : dispersion des voyelles tenues de chaque locuteur sur le plan F1/F2

Le calcul de la surface des triangles vocaliques sujet par sujet (obtenu par la formule de Héron) permet de raffiner cette observation. Les valeurs de surface, résumées à la figure 3, sont légèrement plus importantes dans le groupe témoin que dans le groupe MP ; elles sont par ailleurs beaucoup plus dispersées chez les sujets parkinsoniens que chez les sujets normaux. Ceci suggère donc, de manière générale, une réduction de l'espace articulatoire chez les parkinsoniens, mais également des

conséquences de la maladie très diversifiées en termes de capacités d'exploitation de l'espace vocalique.

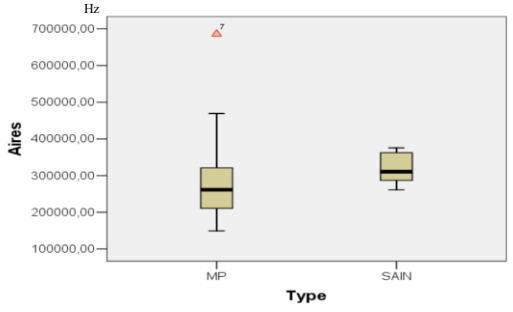


FIGURE 3 : distribution des aires des triangles vocaliques selon le groupe de sujet (moyenne, écart type et écart interquartile).

Dans la majorité des cas, les aires des triangles vocaliques sont inférieures chez les sujets atteints de la MP, mais certains sujets se caractérisent par des valeurs voisines de celles caractérisant les sujets sains (HP et GF), voire supérieure (BD).

3.2 L'amplitude du mouvement

Les comparaisons des distances euclidiennes « aller » et « retour » dans le plan F_1/F_2 mettent en évidence des différences significatives (respectivement, F= 22.728, p<.001, dl= 1 et F= 14.764, p<.001, dl= 1) entre les deux groupes de sujets. La figure 4 illustre ces différences en procédant logatome par logatome.

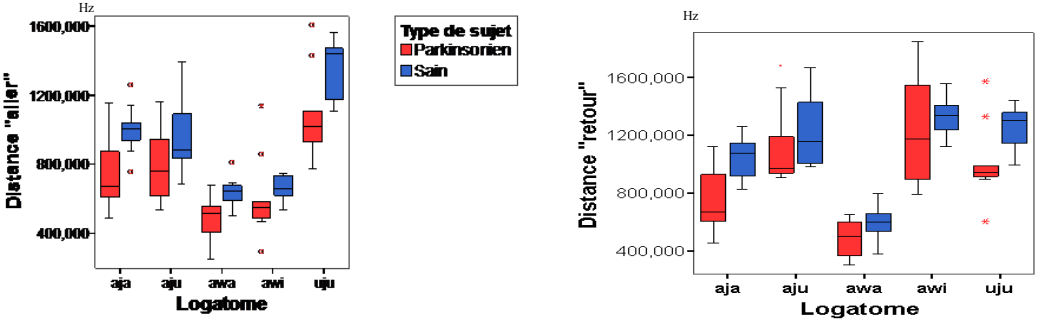


FIGURE 4 : distance « aller » (gauche) et « retour » (droite) des deux groupes de participants selon les logatomes produits

Tant pour la section « aller » que pour la section « retour », la distance est supérieure chez les sujets sains. Ceci s'observe en outre quel que soit le type de glide considéré. En ce qui concerne la distance « aller », les différences sont particulièrement marquées pour les logatomes /aja/ et /uju/. La variabilité au sein du groupe MP est supérieure à celle du groupe contrôle pour /aja/ mais est par contre relativement inférieure pour la production de /uju/. Pour la distance « retour », la variabilité

lors de la production de /awi/ est fortement marquée dans le groupe MP et particulièrement supérieure à celle du groupe contrôle tandis que la variabilité concernant /uju/ est fortement réduite par rapport au groupe contrôle.

3.3 La localisation du glide dans le plan F_1/F_2

Nous avons dans un premier temps questionné la proximité, dans le plan F_1/F_2, des sections initiales et finales des logatomes par rapport aux monophtongues correspondantes (calcul des distances à la référence entre le Départ et V_1 et entre l'Arrivée et V_2, cf 2.4. ci-dessus).

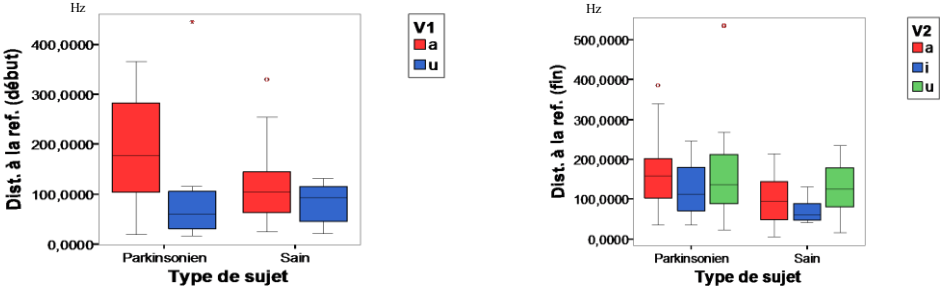


FIGURE 5 : variabilité pour la distance à la référence: D-V_1 (gauche) et A-V_2 (droite)

Cette analyse montre que l'éloignement du point correspondant à la monophtongue est significativement plus grand, tant pour V_1 que pour V_2, dans les logatomes du groupe parkinsonien (respectivement F= 13.964, p<,001, dl= 1 et F= 8.919, p= .004, dl= 1). Ceci pourrait s'interpréter comme relevant d'une stratégie de facilitation par laquelle le sujet parkinsonien produit, au début comme à la fin du logatome, des timbres plus influencés par la cible articulatoire du glide que par la voyelle initiale ou finale du logatome, ce qui permet d'exécuter un mouvement de moindre ampleur que ce n'est le cas chez le locuteur sain.

Cette interprétation appelle cependant une autre question : celle de la résistance du timbre-cible du glide. La réponse est apportée par l'étude de la distance entre la section centrale du glide (au point d'inflexion I) et la cible acoustique de ce dernier, ici matérialisée par la voyelle tenue correspondante (/i/ pour /j/ et /u/ pour /w/). L'analyse montre que dans ce cas précis, il n'y a pas de différence entre le groupe parkinsonien et le groupe de sujets sains. La maladie n'affecte donc pas le comportement du locuteur par rapport à ce qui fait l'identité du glide ; si les parties initiale et finale de ce dernier peuvent se distancier des voyelles correspondantes, il n'en est rien en ce qui concerne le centre-même du glide, qui conserve son identité de la même manière chez le parkinsonien que chez le sujet sain.

4 Conclusions

Notre recherche est partie d'un questionnement : celui de la potentielle informativité des glides pour une meilleure compréhension du contrôle des gestes articulatoires chez les parkinsoniens. Notre intérêt pour ces sons de parole partait d'une réflexion à caractère hypothétique : puisque les glides se caractérisent par une évolution continuelle de leur timbre durant leur production et puisque cette variation acoustique continue est le résultat direct de la variabilité de la géométrie des cavités

résonantielles supra-glottiques, l'analyse de la dynamique acoustique doit très directement renseigner sur les mouvements des articulateurs. Or, comme la MP a pour conséquences des altérations importantes du contrôle des mouvements, il peut paraître sensé d'étudier par ce truchement indirect le contrôle moteur chez les locuteurs porteurs de cette pathologie. Nous avons ainsi observé des parkinsoniens tant en situation de production de monophtongues qu'en situation de production de logatomes de structure VCV. Nous avons comparé leurs comportements phonétiques à ceux de locuteurs sains.

Nous avons observé d'une part une tendance globale à la réduction du champ articulatoire vocalique chez les sujets pathologiques, mais aussi, d'autre part, une variabilité accrue de ce dernier en fonction des individus MP. Nous avons, par ailleurs, constaté, sur base de la comparaison des variations de timbre, que les mouvements articulatoires réalisés lors la production des glides sont de moins grande amplitude pour les sujets atteints de la maladie de Parkinson que pour les sujets sains : les sections initiales et finales des logatomes ressemblent moins, dans le groupe Parkisonien, aux voyelles du système que ce n'est le cas pour les sujets sains. Nous avons par contre observé que la similarité entre la zone médiane du glide et la voyelle cible correspondante ne varie guère en fonction du groupe. Il apparaît donc que si le parkinsonien déploie une stratégie hypoarticulatoire, il le fait en préservant plus le centre du glide que ses extrémités (ce qui, notons-le au passage, n'est pas dépourvu d'intérêt du point de vue de la recherche fondamentale en phonétique).

Il importe de noter que ces observations statistiquement significatives se sont fait jour au départ de productions émises par des sujets assez faiblement atteints sur le plan de l'articulation et qui ne formulent en tout cas guère de plainte à ce sujet; les experts exposés à ces enregistrements n'ont par ailleurs pas détecté de particularité à l'écoute des logatomes produits. Notre analyse se montre donc apte à révéler des phénomènes de nature infra-clinique ; elle permet, par ailleurs d'en offrir une évaluation quantitative, ce qui pourrait être intéressant en perspective du développement d'outils de dépistage précoce, voire de techniques d'accompagnement de la rééducation orthophonique.

Des collectes d'information à plus large échelle doivent maintenant être menées afin de confirmer les éléments mis au jour et surtout de les raffiner en tenant compte plus étroitement des caractéristiques médicales des patients et de l'évaluation des relations entre traitement de la parole et qualité de vie, thématique centrale du projet dans le cadre duquel s'inscrit la présente étude. La prochaine collecte aura également pour but d'accroitre le nombre de productions par sujet et de rendre la liste des logatomes exhaustive, et ce par l'ajout des logatomes /iwi/ ; /uja/ ; /iwa/. Par ailleurs, la mise au point d'indices de la dynamique plus précis nécessitera sans aucun doute des développements mathématiques additionnels.

Même si déjà de nombreuses contributions ont été publiées dans le domaine des répercussions vocales de la maladie de Parkinson, la poursuite de travaux du type de ceux présentés ici nous paraît judicieuse pour plusieurs raisons : d'une part l'originalité de l'approche acoustico-articulatoire que nous tentons de développer, d'autre part, le caractère encore passablement obscur de certains des aspects de la maladie, et en particulier de son étiologie ainsi que de ses manifestations précoces, enfin, l'enjeu de santé publique constitué par une maladie neuro-dégénérative qui est la deuxième en importance, du point de vue épidémiologique, après la démence d'Alzeihmer. Qui plus est, la plupart des études aujourd'hui publiées concernent, de facto, des locuteurs anglophones : pour environ 350 articles recensés sur Medline/PubMed et sur Scopus, seule une trentaine concerne des locuteurs francophones, soit à peine 10% des recherches référencées. La communauté parole de Francophonie a ici un défi à relever.

Références

ACKERMANN, H., HERTRICH, I., & HEHR, T. (1995). Oral diadochokinesis in neurological dysarthrias. *Folia Phoniatr Logo.* 47, 15–23.

ACKERMANN, H., & ZIEGLER, W. (1991). Articulatory deficits in Parkinsonian dysarthria: An acoustic analysis. *J. Neurol Neurosur Ps*, 54, 1093–1098.

BANG, YI., MIN, K., SOHN, YH., & CHO, SR. (2013). Acoustic characteristics of vowel sounds in patients with Parkinson disease. *NeuroRehabilitation, 32*(3), 649-54.

CHEANG, H.S., & PELL, M.D. (2007). An acoustic investigation of Parkinsonian speech in linguistic and emotional contexts. *J Neurolinguist*, 20, 221–241.

COUVREUR, N., BRUYNINCKX, M., & HARMEGNIES, B. (1999). Effects of parkinsonian symptoms on voiced palatals. *Proceedings of 14th International Congress of Phonetic Sciences,1*, 831-834.

DARLEY, F.L., ARONSON, A.E., & BROWN, J.R. (1975). *Motor Speech Disorders*, Philadelphia: W.B. Saunders Company.

DUEZ, D., JANKOWSKI, L., PURSON, A., & VIALLET, F. (2012). Some prosodic characteristics of parkinsonian French speech: Effects of bilateral stimulation of the subthalamic nucleus. *Journal of Neurolinguistics, 25*,104–120

GAMBOA, J., JIMENEZ- JIMENEZ, F.J., NIETO, A., MONTOJO, A., ORTI-PAREJA, M., MOLINA, J.A., *et al.* (1997). Acoustic voice analysis in patients with Parkinson's disease treated with dopaminergic drugs. *J Voice*, 11, 314–320.

GHIO, A., ROBERT, D., GRIGOLI, C., MAS, M., DELOOZE, C., MERCIER, C., & VIALLET, F. (2014). F0 characteristics in Parkinsonian speech: Contrast between the effect of hypodopaminergy due to Parkinson's disease and that of the therapeutic delivery of L-Dopa. *Rev Laryngol Otol Rhinol (Bord), 135*(2), 63-70.

GOBERMANN, A.M., & COELHO, C. (2002). Acoustic analysis of parkinsonian speech I: Speech characteristics and L-Dopa therapy. *NeuroRehabilitation,* 17, 237-246.

HERTRICH, I., LUTZENBERGER, W., SPIEKER, S., & ACKERMANN, H. (1997). Fractal dimension of sustained vowel productions in neurological dysphonias: an acoustic and electroglottographic analysis. *J Acoust Soc Am, 102*(1), 652-654.

LOGEMANN, J. A., & FISCHER, H. B., BOSHES, B. & BLONSKY, E. R. (1978). Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. *J Speech Hear Disord,* 43, 47–57.

SAUVAGEAU, M., ROY, JP., CANTIN, L., PRUD'HOMME, M., LANGLOIS, M., & MACOIR, J. (2015). Articulatory Changes in Vowel Production following STN DBS and Levodopa Intake in Parkinson's Disease. *Parkinson's Disease*, Article ID 382320.

TJADEN, K. (2000). An acoustic study of coarticulation in dysarthric speakers with Parkinson disease. *J Speech Lang Hear Res, 43*(6), 1466-80.

TJADEN, K. & SUSSMAN, J. (2006). Perception of coarticulatory information in normal speech and dysarthria. *J Speech Lang Hear Res, 49*(4), 888-902.

WALSH, B., & SMITH, A. (2012). Basic parameters of articulatory movements and acoustics in individuals with Parkinson's disease. *Mov Disord.* 27, 843-850.

WONG, M.N., MURDOCH, B.E., & WHELAN, B.-M. (2011). Lingual Kinematics in Dysarthric and Nondysarthric Speakers with Parkinson's Disease. *Parkinson's Disease*, Article ID 352838.

Dénomination d'image versus détection interne de phonème : deux méthodes pour étudier la planification de la production de parole

Pierre Hallé [1, 2, 3] Laura Manoiloff [4] Juan Segui [2]
(1) Laboratoire de Phonétique et Phonologie, 75005 Paris, France
(2) Laboratoire Mémoire et Cognition, 92100 Boulogne-Billancourt, France
(3) Haskins laboratories, CT 06511 New Haven, USA
(4) Laboratorio de Psicologia Cognitiva, X5000 Cordoba, Argentina
pierre.halle@univ-paris3.fr, lmvmanoiloff@gmail.com,
juan.segui@parisdescartes.fr

RESUME

Cette étude est motivée initialement par une question méthodologique : la validité des mesures de temps de dénomination d'image, très utilisés pour explorer les processus de planification de production de parole. Le temps de dénomination est le temps écoulé entre affichage de l'image et début acoustique de la réponse verbale. Dans cet article, nous résumons la littérature sur les inconvénients de cette mesure. Nous présentons ensuite notre étude, qui compare directement temps de dénomination d'image et temps de détection interne de phonème initial. Les participants sont hispanophones. Les noms d'image sont contrastés en fréquence lexicale et phonème initial. Les temps de réponse pour les deux mesures sont assez proches. Cependant, ceux de détection de phonème sont relativement insensibles au type de phonème initial, contrairement aux temps de dénomination. Au delà de l'avantage méthodologique de la détection interne de phonème, nos données suggèrent que celle-ci opère sur des représentations relativement abstraites.

ABSTRACT

Picture naming versus internal phoneme monitoring: two methods for exploring speech production planning.
A methodological issue initially motivated the present study: how valid are the measurements of picture-naming times, which are widely used to investigate the processes of speech production planning? Picture-naming time is the time elapsed from picture display to the acoustic onset of verbal response. We begin with a review of the literature pointing to the shortcomings of naming times. We then report our own study, which directly compares picture-naming times and internal monitoring times of word-initial phonemes. Response times for both measurements are quite similar. However, the phoneme detection response times are relatively insensitive to word-initial phoneme's phonetic type, contrary to picture-naming times. Beyond the methodological advantage of internal phoneme monitoring over naming, our data suggest that internal monitoring bears on rather abstract internal representations.

MOTS-CLES : production de parole, dénomination d'image, détection interne de phonème.
KEYWORDS: speech production planning, picture-naming, internal phoneme monitoring.

1 Introduction

La tâche de dénomination d'image a une très longue histoire (Cattell, 1885). Elle a été et reste encore très utilisée pour l'étude de la production de parole. Dans cette tâche, on présente au sujet une série d'images, une par une, que le sujet doit nommer aussi rapidement que possible. La variable dépendante est le "temps de dénomination" mesuré entre l'apparition de l'image et le début de la réponse vocale. De nombreuses études ont utilisé un dispositif appelé "clef vocale" pour effectuer cette mesure. La clef vocale se "déclenche" au début de la réponse vocale, mesurant alors un temps de réponse. Il est naturellement possible que le sujet produise une réponse "incorrecte" par exemple *chat* pour l'image d'un chien, voire un hypéronyme non attendu comme *quadrupède*, *animal*, etc. pour le même chien, voire encore qu'il ne réponde pas mais que la clef se déclenche malgré tout, par exemple sur un bruit de toux. Il est donc nécessaire de vérifier les réponses en temps réel ou bien d'enregistrer les productions vocales des sujets et vérifier après l'expérience si les mesures de la clef vocale sont acceptables ou non.

Même après ces vérifications, les valeurs données par les clefs vocales ne sont pas entièrement satisfaisantes. Rastle et Davis (2002) ont montré que les deux types existants de clef vocale (à seuil simple ou à intégration) se déclenchent systématiquement *après* le départ acoustique. Il ressort clairement de cette étude que la façon de mesurer les temps de dénomination la plus précise est d'enregistrer les réponses vocales et de repérer le départ acoustique de l'énoncé produit par le sujet —s'il s'agit d'une réponse correcte— à partir du signal acoustique et son spectrogramme.

Même en appliquant cette procédure quelque peu coûteuse (mais qui peut être assistée de façon semi-automatique par un algorithme de détection de départ acoustique), un problème subsiste lorsque l'on souhaite comparer des mots commençant par des consonnes différentes. En particulier, la variation de temps de dénomination induite par les différences entre consonnes initiales a été examinée par Rastle et al. (2005). Ces auteurs utilisent une tâche de dénomination retardée de transcriptions phonétiques (les sujets ont au moins 5 ans de formation en phonétique). Les sujets ne prononcent la transcription affichée sur un écran que lorsqu'un ton bref a été émis. Cette procédure permet de mesurer des temps de dénomination une fois que la phase de préparation de la production est achevée : seules des différences liées à l'articulation des réponses verbales peuvent influencer les temps. Rastle et al. observent que des différences phonétiques entre phonèmes initiaux induisent une variation substantielle et suggèrent aux chercheurs qui utilisent les temps de dénomination de se limiter à la comparaison de mots cibles partageant le même phonème initial. Nous avons repris leurs données dans la figure 1 qui montre clairement que les temps les plus longs sont systématiquement associés aux occlusives, en particulier les /b, d, g/ anglais réalisés [p, t, k]

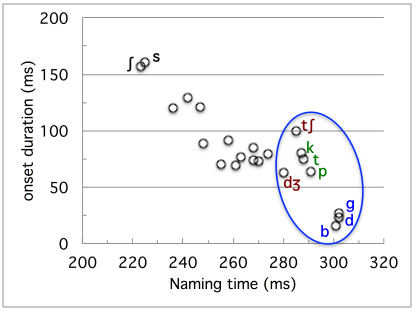


FIGURE 1 : temps de dénomination x durée acoustique du phonème initial, par phonème.

Nous nous plaçons ici dans le cadre du modèle de Levelt (e.g., Levelt et al., 1999). Dans ce modèle, l'accès lexical, par exemple au nom d'une image, fournit une représentation phonologique abstraite consistant d'une part en une séquence de phonèmes et d'autre part en un cadre métrique et prosodique. Ces informations sont combinées, éventuellement en fonction des mots adjacents, par des opérations d'encodage phonologique dont le résultat est une spécification phonologique complète intégrant la métrique, et en particulier la syllabation. Levelt considère cette spécification syllabée comme le point de départ de la construction d'un programme articulatoire (qu'il appelle 'phonetic plan') constitué de 'gestural scores' syllabiques. Dans ce cadre de travail, le temps de réponse que l'on voudrait idéalement mesurer est celui de la sortie de l'encodage phonologique, point de départ pour lancer l'articulation. Le plus proche de ce temps idéal est le départ *articulatoire* de la réponse vocale, modulo un décalage que l'on peut supposer relativement stable, bien qu'il varie sans doute en fonction des différents ensembles d'articulateurs engagés dans l'articulation physique et donc des différents paramètres physiologiques d'inertie et de contrôle musculaire qui leur sont associés.

Les données de Rastle et al. (2005) montrent que le temps de dénomination s'éloigne nettement de cet idéal, même lorsqu'il est mesuré avec précision, entre le début de l'affichage de l'image et le début acoustique de la réponse vocale. C'est en particulier le cas lorsque le mot produit commence par une occlusive non-voisée. Pour ces occlusives, le départ articulatoire est le début du geste d'occlusion. Il précède le début acoustique qui ne peut-être que le relâchement de l'occlusion. On ne peut observer la trace acoustique du début de l'occlusion que dans le cas des occlusives pré-voisées : c'est approximativement le début du pré-voisement. Il n'est donc pas étonnant que Rastle et al. (2005) trouvent des temps de dénomination plus longs pour les occlusives —y compris les affriquées— que les sonantes et les fricatives, puisque toutes les occlusives en position initiale sont non-voisées phonétiquement en anglais. C'est bien ce que montre la figure 1.

Cependant, mesurer le départ articulatoire requiert un appareillage spécialisé et n'est de toute façon pas une chose facile à mettre en œuvre. Schaeffler et al. (2014) ont comparé les temps de réponse de dénomination d'image mesurés à partir de l'articulation (enregistrements ultra-son de la langue et vidéo des lèvres) vs. à partir du signal acoustique. Ils trouvent que le départ articulatoire précède le départ acoustique d'environ 120 à 180 ms. Autrement dit, la parole audible est précédée par des mouvements silencieux des articulateurs. Schaeffler et al. (2014) trouvent aussi que le lieu (labial, lingual) et la manière (voyelle, occlusive, fricative) d'articulation du phonème initial n'ont pas d'effet sur les temps mesurés pour les départs articulatoires. Donc, en particulier, pas de différence entre occlusives et fricatives. Ces mesures se rapprochent donc de la mesure idéale mentionnée plus haut mais au prix d'une lourdeur technique prohibitive pour les études de production de parole.

Dans cette étude, exploratoire par sa dimension, nous comparons temps de dénomination d'image et temps de détection interne du phonème initial du nom de l'image avec le même matériel expérimental, c'est à dire les mêmes images. Nous tentons d'abord de répondre à la question suivante : Quelle est la sensibilité des deux mesures de temps de réponse aux variations liées aux différents types de phonème initial ? L'étude de Rastle et al. (2005) suggère que, toutes choses égales par ailleurs, les temps de dénomination mesurés à partir du signal acoustique sont plus longs pour les occlusives que les autres types de phonème. L'étude de Schaeffler et al. (2014) suggère que le type de phonème initial, par exemple occlusive vs. fricative, n'a pas d'effet sur les temps de dénomination mesurés à partir de l'articulation. En est-il de même pour les temps de détection interne de phonème initial ? Un objectif secondaire de cette étude est de comparer les effets bien connus de fréquence lexicale en production (Oldfield et Wingfield, 1965) sur les temps de dénomination et les temps de détection interne. L'effet de fréquence lexicale sur les temps de

détection interne de phonème initial a été mis en évidence par Manoiloff et al. (2013). Cependant, ces effets n'ont encore jamais été comparés avec ceux obtenus pour les temps de dénomination : sont-ils comparables pour les deux mesures ? Ou bien l'une des deux est-elle plus sensible que l'autre ? Pour répondre à cette question ainsi qu'à la question principale sur les effets du type de phonème initial, nous avons repris le matériel utilisé par Manoiloff et al. (2013). Ceci permet en même temps de vérifier si les effets de fréquence lexicale en détection interne de phonème initial peuvent ou non être répliqués.

2 Expériences : dénomination vs. détection interne de phonème

Nous avons testé deux groupes d'une trentaine de sujets argentins hispanophones sur les mêmes images. Pour un groupe, les sujets devaient dénommer les images présentées. Pour l'autre groupe, les sujets devaient détecter des phonèmes pré-spécifiés dans le nom des mêmes images.

2.1 Méthode

– Participants

Soixante-et-un étudiants en Psychologie à l'Université Nationale de Cordoba (Argentine), âgés entre 18 et 26 ans, ont participé volontairement aux expériences. Trente ont participé à l'expérience de détection de phonème et 31 autres à l'expérience de dénomination d'image. Tous étaient locuteurs natifs de l'espagnol parlé en Argentine. Aucun d'entre eux ne souffrait de déficit de vision ou de trouble du langage.

– Matériel

Nous avons repris le matériel utilisé par Manoiloff et al. (2013). Ce matériel consiste en 30 dessins expérimentaux en noir et blanc, sélectionnés parmi les 400 dessins de l'étude de Cycowicz et al. (1997), qui inclut les 260 dessins de Snodgrass et Vanderwart (1980). Quinze dessins ont un nom de haute fréquence et 15 un nom de basse fréquence, selon les tables de fréquence établies par Alameda et Cuetos (1995) et selon le jugement subjectif de trente sujets n'ayant pas participé à l'étude. La table 1 montre les caractéristiques de ces trente noms d'image. Le pourcentage d'accord sur le nom des images et leur complexité visuelle sont tirés des normes de Manoiloff et al. (2010). Il est à noter que les pourcentages d'accord sur le nom sont très élevés. À ces 30 items expérimentaux s'ajoutent 227 items de remplissage pour la phase de test et 18 ou 20 items pour une phase d'entraînement familiarisant les sujets à l'une des deux tâches. Ces items supplémentaires étaient extraits du matériel utilisé par Cycowicz et al. (1997) et Manoiloff et al. (2010).

caractéristiques	noms HF	noms BF
fréquence des tables de Alameda et Cuetos (o.p.m.)	138.1	5.3
fréquence subjective (échelle 1-5)	4.1	1.6
nombre de syllabes	2.3	2.5
nombre de phonèmes	5.1	5.8
pourcentage d'accord sur le nom	98.8%	96.7%
complexité visuelle (échelle 1-5)	2.6	3.0

TABLE 1 : Caractéristiques du matériel : noms de haute vs. basse fréquence d'usage (HF vs. BF).

Bien que ce matériel n'ait pas été conçu pour tester l'effet possible du type de phonème initial sur la dénomination ou la détection interne, il se trouve que les types principaux qu'il est intéressant de tester sont bien représentés, même si de façon non-homogène. Parmi les 30 noms d'image, 17 commencent par une occlusive sourde et 12 par une consonne non-occlusive. Le mot *gato* "chat" est le seul représentant de la catégorie occlusive sonore : pour cette raison et pour d'autres raisons qui sont exposées par la suite, nous n'avons pas retenu les données pour cet item. La table 2 montre les différentes consonnes initiales des 30 noms d'image utilisés.

Consonne	l	r	m	f	g	p	t
Compte	2	4	3	3	1	12	5

TABLE 2 : Distribution des consonnes initiales pour les 30 noms d'image.

– Design

Pour l'expérience de détection de phonème, les items étaient présentés en 7 blocs correspondant chacun à l'un des 7 phonèmes à détecter. Ces blocs comprenaient entre 25 et 72 items : les items test (entre 1 et 12 : cf. Table 2), des items de remplissage avec le phonème initial des items test (entre 0 et 4), et des items de remplissage avec un phonème initial différent (entre 20 et 60 de sorte à homogénéiser la proportion d'items test). L'ordre de présentation des blocs, et celui des items à l'intérieur des blocs, était pseudo-aléatoire et différent pour chaque sujet.

Pour l'expérience de dénomination, les mêmes items (au total, 257 = 30 + 227) étaient présentés aux sujets dans un ordre pseudo-aléatoire différent pour chaque sujet.

– Procédure

Pour la détection de phonème, l'expérimentateur expliquait aux sujets qu'on allait leur présenter des jeux d'images, une image à chaque essai, et qu'ils devaient répondre le plus rapidement possible en appuyant sur le bouton-réponse aux images dont le nom commençait par le phonème-cible spécifié pour le jeu en cours. Pour chacun des sept blocs (i.e., des sept "jeux") d'images, l'expérimentateur indiquait oralement (en espagnol) le phonème à détecter par des exemples (e.g., "maintenant vous devez répondre aux images dont le nom commence par le son /t/ comme dans *tambour*, *taureau*, etc."). Il était bien précisé au sujet qu'il devait répondre à un son et pas à une lettre. L'expérience proprement dite (la phase de test) était précédée par une phase d'entraînement comprenant six blocs de trois essais, chacun associé à un phonème-cible non utilisé dans la phase de test.

Pour la dénomination, il était expliqué aux sujets qu'ils devaient dénommer l'image présentée à chaque essai aussi rapidement que possible. Les sujets portaient un micro-casque pour éviter une trop grande variabilité de l'intensité sonore. Les réponses vocales étaient directement numérisées et enregistrées (le temps 'zéro' correspondant à l'apparition de l'image à dénommer) pour repérage ultérieur du départ acoustique des mots produits par les sujets. Le dispositif comprenait aussi une estimation logicielle du temps de réponse (à partir du signal acoustique enregistré). Nous n'avons pas analysé les temps donnés par cette estimation. La phase de test était précédée par une phase d'entraînement comportant 20 images non utilisées dans la phase de test.

Pour les deux expériences, les images étaient présentées sur écran d'ordinateur en noir sur fond blanc dans un carré de 7 cm de côté au centre de l'écran. Chaque image restait affichée jusqu'à la réponse du sujet (indiquée par le déclenchement du bouton réponse ou par la clef vocale) ou bien pendant deux secondes en cas d'absence de réponse. L'essai suivant était initié une seconde après la

disparition de l'image. Pour les deux expériences, les phases d'entraînement et de test étaient précédées par une phase de "familiarisation" où les 30 images expérimentales étaient présentées une par une, avec le nom attendu affiché en dessous, pendant 2 s avec un intervalle de 1 s entre deux images. Aucune réponse n'était demandée aux sujets. Cette phase était destinée à minimiser les réponses inattendues malgré le pourcentage élevé d'accord sur le nom des images (>97%). L'expérience était contrôlée avec le logiciel DMDX (Forster et Forster, 2003) et durait environ 40 minutes (détection) ou 30 minutes (dénomination).

2.2 Résultats

– Données retenues

Un des 31 sujets de l'expérience de dénomination a été exclu des analyses car les réponses vocales enregistrées étaient beaucoup trop bruitées. L'item gato ("chat") a été exclu des analyses car il a été produit 11 fois sur 30 avec une fricative proche de [ɣ] au lieu de l'occlusive pré-voisée [g]. Les données manquantes de dénomination (pas de réponse avant 2 s ou réponse erronée : e.g., *ganzo* "oie" pour *pato* "canard") représentent 2.5% des 870 essais retenus (29 items x 30 sujets). Les données manquantes de détection (pas de réponse) représentent 3.9% des essais. Nous avons aussi exclu les temps de réponse supérieurs à 1500 ms. Ce qui élimine 2.5% des données de détection mais aucune donnée de dénomination.

– Effet de fréquence lexicale

La figure 2 montre les temps de réponse (TR) de dénomination et de détection. L'effet de fréquence est manifeste pour les deux tâches. Nous avons conduit des analyses de variance par sujets (F1) et par items (F2) avec TR comme variable dépendante, la Fréquence lexicale (haute vs. basse) comme facteur inter-item et intra-sujet et la Tâche (dénomination vs. détection) comme facteur intra-item et inter-sujet. Le facteur Fréquence est très significatif globalement, $F1(1,58)=163.43$, $p<.00001$; $F2(1,27)=25.34$, $p<.00001$, mais aussi pour chacune des deux tâches, $p1s<.00001$, $p2s<.001$. Le facteur Tâche n'est pas significatif, $F1(1,58)=1.76$, $p=.19$; $F2(1,27)=2.19$, $p=.15$. Cependant, l'interaction Tâche x Fréquence, marginalement significative par sujets, $F1(1,58)=3.42$, $p=.069$, suggère un effet de fréquence plus grand pour la dénomination que la détection (121 vs. 90 ms). Cette différence n'est cependant ni robuste ni numériquement impressionnante : l'effet de fréquence lexicale est de l'ordre de 100 ms pour les deux tâches, ce qui est proche des résultats de Manoiloff et al. (2013). De plus, les TRs de détection de phonème ne sont pas significativement plus longs, dans l'ensemble, que les TRs de dénomination.

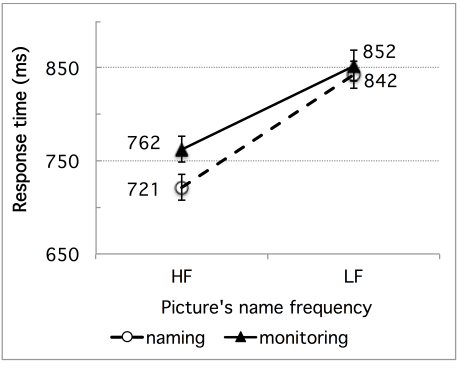


FIGURE 2 : TRs en fonction de la tâche (dénomination vs. détection) et de la fréquence.

– Effet du type de phonème initial

Nous avons regroupé les phonèmes retenus en deux catégories : occlusives non-voisées (/p, t/) et autres (/f, m, l, r/). Comme le montre la figure 3A, les temps de réponses pour ces deux catégories sont en moyenne pratiquement les mêmes pour la détection interne, alors qu'ils sont nettement plus longs avec les occlusives que les non-occlusives pour la dénomination. Cette observation est confirmée par deux analyses de variance des temps de réponse, par sujets et par items, avec la Tâche (dénomination vs. détection) comme facteur intra-item et inter-sujet, et le type de Consonne initiale (occlusive vs. non-occlusive) comme facteur inter-item et intra-sujet. L'interaction entre ces deux facteurs n'est pas significative dans l'analyse par items mais l'est dans celle par sujets, $F1(1,58)=7.63$, $p<.01$, indiquant un effet du facteur Consonne significatif pour la dénomination, $F1(1,29)=33.25$, $p<.00001$; $F2(1,27)=3.34$, $p=.079$ (marginal), mais pas la détection, $F1(1,29)=2.78$, $p=.11$; $F2(1,27)=1.03$, $p=.32$. Ceci suggère que la tâche de détection est moins sensible que celle de dénomination aux variations du type de phonème initial du mot cible. Mais cette analyse ne tient cependant pas compte de l'effet de fréquence lexicale qui, comme le montrent la figure 2 et les statistiques correspondantes, est extrêmement robuste. Il est en principe possible que l'effet du type de phonème initial que nous trouvons (le plus clairement dans l'analyse par sujets) soit confondu en partie avec l'effet de fréquence, bien que les items expérimentaux aient été répartis à peu près équitablement dans les deux sous-ensembles haute et basse fréquence. À tout le moins, la fréquence lexicale peut perturber voire fausser la comparaison entre occlusives et non-occlusives. Des analyses de covariance, avec la fréquence comme covariable, sont donc nécessaires pour mesurer les effets possibles du type de phonème initial indépendamment de la fréquence. Nous avons conduit deux telles analyses de covariance, par sujets et par items, avec la fréquence subjective des noms d'image comme covariable. Nous trouvons que l'effet du facteur Consonne (occlusive vs. non-occlusive) est significatif pour les temps de dénomination, $F1(1,177)=17.79$, $p<.00001$; $F2(1,26)=3.26$, $p=.083$ (marginal), mais pas pour les temps de détection, $F1$ et $F2 <1$, ce qui confirme l'analyse précédente. La covariation des temps de réponse avec la fréquence subjective et le phonème initial est illustrée par la figure 3B où les six phonèmes initiaux /l, m, t, p, r, f/ sont classés par ordre décroissant de fréquence subjective. La figure suggère une relation proche de la linéarité entre TR de détection et fréquence subjective, à l'exception des items en /f/. La relation est moins linéaire pour les TRs de dénomination, systématiquement plus courts que ceux de détection pour tous les phonèmes autres que /p/ et /t/, y compris /f/.

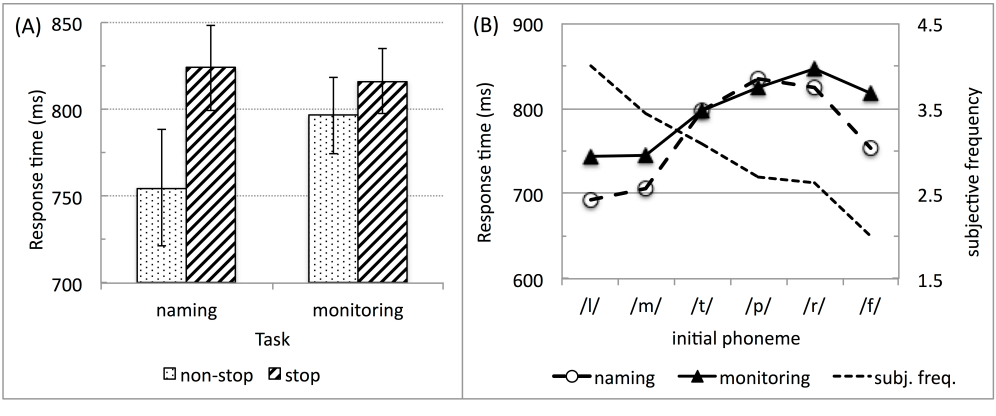


FIGURE 3 : (A) TRs en fonction de la tâche et du type de phonème initial (occlusive ou non) ; (B) TRs détaillés selon les 6 phonèmes utilisés, en ordre décroissant de fréquence subjective.

3 Discussion

Dans cette étude comparant temps de dénomination et temps de détection de phonème initial sur le même ensemble d'images, celui utilisé par Manoiloff et al. (2013), nous répliquons l'effet de fréquence lexicale (Manoiloff et al., 2013) pour les deux types de temps de réponse. L'effet est un peu plus grand pour la dénomination que pour la détection (121 vs. 90 ms), mais la différence n'est pas significative. Cependant, la question principale à laquelle cette étude tente de donner des éléments de réponse est celle de la sensibilité comparée des deux mesures de temps de réponse au type de phonème initial des noms d'image à produire, soit ouvertement (dénomination), soit intérieurement (monitoring interne). Nos données suggèrent que les temps de dénomination sont systématiquement plus longs lorsque les noms d'image commencent par une occlusive que par un autre type de consonne. Les temps de détection ne montrent pas cette différence. Ce résultat est indépendant des effets de fréquence lexicale, comme le montrent nos analyses de covariance.

Une première conséquence de nos résultats, sous réserve qu'ils soient robustes et reproductibles pour d'autres langues que l'espagnol, est un avantage méthodologique du monitoring interne sur la procédure classique de dénomination. En effet, il devient possible de comparer des ensembles de noms d'image sans qu'ils commencent nécessairement par le même type de phonème, puisque les temps de détection interne de phonème initial dans le nom d'une image semblent peu sensibles au type phonétique de phonème. En particulier, les temps de détection, contrairement aux temps de dénomination, ne sont pas plus longs, à fréquence lexicale égale, pour les occlusives que les fricatives, nasales ou liquides.

Au delà de cet avantage méthodologique en faveur du monitoring interne, nos données ont des implications pour le niveau de représentation sur lequel opère la détection interne ou plus largement, le monitoring interne en production de parole. Elles suggèrent que les représentations en jeu sont relativement abstraites. Dans le cadre du modèle de Levelt (Levelt et al., 1999), ces représentations pourraient être celles qui sont récupérées lors de l'accès aux informations lexicales, celles qui sont élaborées au cours de l'encodage phonologique, ou bien celles qui sont produites en sortie de l'encodage phonologique. La dernière possibilité est plus compatible avec la vision de l'encodage phonologique comme un module imperméable à l'inspection consciente : seule sa sortie est accessible, pas ses calculs internes. Une autre possibilité est que les phonèmes soient détectés par le système de compréhension de la parole alimenté par la sortie de l'encodage phonologique (ce que Levelt appelle la "boucle interne" 'internal loop'). Les effets de point d'unicité trouvés par Özdemir et al. (2007) vont dans ce sens : ces effets sont en effet typiques du traitement de la parole externe et sont liés à l'accès lexical en modalité auditive. Or, dans une situation de monitoring interne, l'accès lexical a déjà été effectué, avant l'encodage phonologique. Ce qui suggère que la détection ne se fait pas sur la sortie de l'encodeur mais se fait par la boucle interne. D'autres données récentes remettent en question la communauté des traitements de la parole externe et interne (Gauvin et al., 2016). Nos propres données ne permettent pas de trancher la question de savoir si le monitoring interne peut passer par d'autres voies que le système de compréhension/perception de la parole qui serait commun à la parole externe et à la parole interne. Elles suggèrent simplement que le monitoring interne examine des représentations abstraites indépendantes de la spécification du détail phonétique de réalisation des mots.

Références

ALAMEDA J., CUETOS F. (1995). *Diccionario de frecuencias de las unidades lingüísticas del castellano*. Oviedo: Servicio de las publicaciones de la Universidad de Oviedo.

CATTELL J. M. (1885) Über die Zeit der Erkennung und Benennung von Schriftzeichen, Bilden und Farben. *Philosophische Studien* 2, 635-650.

CYCOWICZ Y., FRIEDMAN D., ROTHSTEIN M., SNODGRASS J. (1997). Picture naming by young children: Norms of name agreement, familiarity, and visual complexity. *Journal of Experimental Child Psychology* 65, 171-237.

FORSTER K., FORSTER J. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments and Computers* 35, 116-124.

GAUVIN H., DE BAENE W., BRASS M., HARTSUIKER R. (2016). Conflict monitoring in speech processing: An fMRI study of error detection in speech production and perception. *NeuroImage* 126, 96-105.

LEVELT W., ROELOFS A., MEYER A. (1999). A theory of lexical access in speech production. *Behavioral and Brain Science* 22, 1-75.

MANOILOFF L., ARSTEIN M., CANAVOSO M., FERNANDEZ L., SEGUI J. (2010). Expanded Norms for 400 Experimental Pictures in an Argentinian Spanish-Speaking Population. *Behavior Research Methods*, *Instruments and Computers* 42, 452-460.

MANOILOFF L., SEGUI J., HALLÉ P. (2013). L'effet de fréquence dans l'accès aux propriétés phonologiques des noms d'objets. *L'Année Psychologique* 113, 335-348.

OLDFIELD R., WINGFIELD A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology* 17, 273-281.

ÖZDEMIR R., ROELOFS A., LEVELT W. (2007). Perceptual uniqueness point effect in monitoring internal speech. *Cognition* 105, 457-465.

RASTLE K., DAVIS M. (2002). On the complexities of measuring naming. *Journal of Experimental Psychology: Human Perception and Performance* 28, 307-3.

RASTLE K., CROOT K., HARRINGTON J., COLTHEART M. (2005). Characterizing the motor stage of speech production: Consonantal effects on delayed naming latency and onset duration. *Journal of Experimental Psychology: Human Perception and Performance* 31, 1083-1095.

SCHAEFFLER S., SCOBBIE J., SCHAEFFLER F. (2014). Measuring reaction times: vocalization vs. articulation. Actes de *10th ISSP*, 379-382.

SNODGRASS J., VANDERWART M (1980). A Standardized set of 260 pictures : Norms for name agreement, image agreement, familiarity and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory* 6, 174-215.

Détection automatique d'anomalies sur deux styles de parole dysarthrique: parole lue vs spontanée

Imed Laaridh[1, 2] Corinne Fredouille[1] Christine Meunier[2]
(1) Université d'Avignon, CERI/LIA, Avignon, France
(2) Université d'Aix Marseille, CNRS, LPL UMR 7309, 13100, Aix-en-Provence, France
imed.laaridh@alumni.univ-avignon.fr, corinne.fredouille@univ-avignon.fr,
christine.meunier@lpl-aix.fr

RÉSUMÉ

L'évaluation perceptive de la parole pathologique reste le standard dans la pratique clinique pour le diagnostic et le suivi des patients. De telles méthodes incluent plusieurs tâches telles que la lecture, la parole spontanée, le chant, les mots isolés, la voyelle tenue, etc. Dans ce contexte, les outils de traitement automatique de la parole ont montré leur pertinence dans l'évaluation de la qualité de parole ainsi que dans le cadre de la communication améliorée et alternative (CAA) pour les patients atteints de troubles de parole. Cependant, peu de travaux ont étudié l'utilisation de ces outils sur la parole spontanée. Ce papier examine le comportement d'un système de détection automatique d'anomalies au niveau phonème face à la parole dysarthrique lue et spontanée. Le comportement du système révèle une variabilité inter-pathologique à travers les styles de parole.

ABSTRACT

Automatic anomaly detection for dysarthria across two speech styles : read vs spontaneous speech.

Perceptive evaluation of speech disorders is still the standard method in clinical practice for the diagnosing and the following of the condition progression of patients. Such methods include different tasks such as read speech, spontaneous speech, isolated words, sustained vowels, etc. In this context, automatic speech processing tools have proven pertinence in speech quality evaluation and assistive technology-based applications. Though, a very few studies have investigated the use of automatic tools on spontaneous speech. This paper investigates the behavior of an automatic phone-based anomaly detection system when applied on read and spontaneous French dysarthric speech. The behavior of the automatic tool reveals inter-pathology differences across speech styles.

MOTS-CLÉS : Dysarthrie, traitement automatique de la parole, détection d'anomalies, parole lue, parole spontanée.

KEYWORDS: Dysarthria, automatic speech processing, anomaly detection, read speech, spontaneous speech.

1 Introduction

La dysarthrie est un trouble de la réalisation motrice de la parole. Elle peut avoir comme origine des lésions du système nerveux central et/ou périphérique et affecter différents niveaux de la production de la parole (respiratoire, laryngé, articulatoire, prosodique, etc.). Ces troubles peuvent alors se manifester sous forme de faiblesse, incoordination et mouvements involontaires selon la localisation du ou des

lésions neurologiques. La parole dysarthrique a été étudiée selon différents axes : l'évaluation perceptive de la parole pour la classification de la dysarthrie (Darley *et al.*, 1969; Duffy, 2005; Darley *et al.*, 1975), l'étude des différentes tâches de production de parole (Van Lancker Sidtis *et al.*, 2012; Kempler & Van Lancker, 2002), l'évaluation perceptive de la sévérité de la dysarthrie dans l'étude notamment de l'intelligibilité du locuteur (Enderby, 1983; Yorkston *et al.*, 1996; Lowit & Kent, 2010) et les analyses acoustiques pour caractériser les effets de la dysarthrie sur le signal de parole (Kent *et al.*, 1999; Green *et al.*, 2013). Ces études ont pour objectif d'aider les cliniciens dans leur analyse des altérations de la parole et leur évaluation clinique, cruciale pour le suivi des patients dans le cadre de traitement thérapeutique et/ou de rééducation. Dans ce cadre, le traitement automatique de la parole a été considéré comme une solution potentielle pouvant apporter des outils d'évaluation objectifs des troubles de parole. (Carmichael, 2007; Middag *et al.*, 2009; Kim & Kim, 2012). Par ailleurs, des approches reposant sur ces traitements automatiques telles que la reconnaissance automatique de la parole sont expérimentées dans le cadre d'outils de communication améliorée et alternative (CAA) afin d'accompagner des patients souffrant de troubles de parole dans leur vie quotidienne (Christensen *et al.*, 2013; Parker *et al.*, 2006).

Dans la littérature, des paramètres acoustiques ou perceptifs tels que l'imprécision des consonnes, la distorsion des voyelles, le débit faible et l'hypernasalité sont étudiés pour caractériser les perturbations principales des différents types de dysarthrie au niveau de la production de parole. Cependant, des analyses acoustiques et phonétiques plus précises restent nécessaires afin de considérer la variabilité d'altérations de parole inter- et intra-groupe pathologique (Tomik & Guiloff, 2010) ou de différents styles de parole (lue, spontanée, chantée, mots isolés, etc.). Par ailleurs, pour que de telles analyses soient pertinentes, elles nécessitent un nombre important de patients, une variété de pathologies exhibant différentes dysarthries (spastique, flasque, ataxique, hyper ou hypokinétique, ou mixte) et différents degrés de sévérité afin d'observer leurs effets sur la production de la parole ainsi que les possibles stratégies de compensation établies. Dans ce cadre, les outils de traitement automatique de la parole peuvent apporter des solutions et une aide aux experts humains en focalisant leurs attentions sur des segments et des zones spécifiques de parole (parmi une grande quantité de productions) présentant des caractéristiques acoustiques inattendues par rapport à une production normale.

Dans (Chandola *et al.*, 2007) et dans un contexte plus général, la détection d'anomalies se rapporte au problème de trouver, dans un corpus, des jeux de données qui ne sont pas conformes au comportement typique attendu. Dans le cadre de la parole dysarthrique, les anomalies peuvent être des segments de parole présentant des caractéristiques acoustiques inattendues. Ces segments peuvent avoir différents niveaux de granularité tels que la syllabe ou le phonème. Dans des travaux précédents (Fredouille & Pouchoulin, 2011; Laaridh *et al.*, 2015a), les auteurs ont proposé une approche de détection automatique des anomalies au niveau phonème reposant sur deux phases : un alignement automatique contraint par le texte permettant une segmentation du signal au niveau phonème et une classification de ces segments en phonème normal et anormal (anomalie). De plus, dans (Laaridh *et al.*, 2015b), une étude sur l'alignement automatique de la parole dysarthrique a montré une dépendance entre sa qualité et les pathologies des patients, les catégories phonétiques et la sévérité de la dysarthrie.

Dans cet article [1], les auteurs examinent l'impact du style de parole (lue et spontanée) sur la détection automatique d'anomalies et la phase de classification. En effet, des études comparatives des troubles moteurs de la parole ont trouvé des caractéristiques différentes d'articulation, de débit et de pauses de respiration selon les styles de parole (Brown & Docherty, 1995). En plus, il est possible que les locuteurs dysarthriques développent des stratégies afin d'éviter des contextes linguistiques "difficiles".

1. Ce travail est soutenu par le Labex BLRI (ANR-11-LABEX-0036), le projet A*MIDEX (ANR-11-IDEX-0001-02) financé par le programme "investissements d'avenir" du gouvernement Français géré par l'ANR et le projet Typaloc (ANR-12-BSH2-0003-03).

De telles stratégies ne peuvent être appliquées que dans le cadre de parole spontanée. Dans ce contexte, il est intéressant d'étudier si notre système de classification est capable de détecter des anomalies sur de la parole spontanée et s'il présente le même comportement que face à la parole dysarthrique lue.
Le reste de cet article est structuré comme suit. La section 2 décrit l'approche de détection automatique d'anomalies utilisée. Les corpus de données utilisés sont présentés dans la section 3. Dans la section 4, le comportement du système face aux styles de parole différents est analysé. La section 5 reprend quelques conclusions et perspectives de travail.

2 Détection automatique d'anomalies

L'approche de détection d'anomalies étudiée repose sur deux phases. La première est un alignement contraint par le texte. La deuxième est une classification supervisée des phonèmes en deux classes (normal et anormal). Dans chaque classe, les phonèmes sont caractérisés par un ensemble de paramètres jugés pertinents pour la tache de classification.

2.1 Alignement automatique au niveau phonème

L'alignement des enregistrements de parole en phonèmes est réalisé grâce à un outil d'alignement automatique contraint par le texte. Cet outil utilise comme entrées la séquence de mots prononcée dans chaque enregistrement ainsi qu'un lexique phonétisé présentant une variété phonologique de chaque mot basée sur un ensemble de 37 phonèmes de la langue française. La séquence de mots est le résultat d'une transcription manuelle réalisée par un humain suivant un ensemble de règles d'annotation pour les ajouts, suppressions et substitutions. Lors de cette transcription, des unités inter-pausales (UIP) sont aussi annotées. Une UIP est définie comme une unité de parole ne contenant pas de pauses et séparée d'autres UIP par au moins 250ms de silence.
L'alignement automatique repose sur un décodage du signal de parole par l'algorithme Viterbi basé sur des modèles statistiques (des modèles de Markov cachés, HMM) associés à chaque phonème. Dans ce travail, chaque phonème est représenté par un HMM à trois états, indépendant du contexte construit par estimation du maximum de vraisemblance sur la base d'environ 200 heures d'enregistrements radiophoniques français (Galliano *et al.*, 2005). Une adaptation de type Maximum A Posteriori (MAP) à 3 itérations est appliquée afin de créer des modèles dépendants du locuteur. Ce processus d'alignement résulte en une segmentation temporelle des enregistrements en phonèmes avec, pour chaque phonème, ses frontières de début et de fin dans le signal.
La qualité de cet alignement automatique par rapport à un alignement de référence (manuel) a fait l'objet d'une étude dans (Laaridh *et al.*, 2015b) qui a montré une dépendance entre qualité de l'alignement et sévérité de la dysarthrie des patients. De plus, une large variabilité inter-pathologique et inter-phonémique de la précision de l'alignement a été relevée.

2.2 Classification de la parole

Cette étape permet la caractérisation de chaque phonème par un ensemble de paramètres essentiellement tirés de l'alignement automatique de la parole. Pour chaque phonème p et son segment associé y_p, les paramètres suivant sont extraits :

— la durée du segment y_p associé au phonème p, exprimée en terme de nombre de trames de 10ms ;
— le nombre de trames de y_p pour lesquelles la recherche du meilleur état au sein des modèles HMM conduit à un état du phonème p ;

— le score acoustique du phonème p' émanant de la recherche du meilleur modèle HMM sur la base du segment y_p ; si p est le meilleur phonème, le second dans l'ordre est considéré à sa place ;
— la catégorie phonétique de p' ;
— le score acoustique du second meilleur phonème p'' correspondant au segment y_p ; si p est l'un des deux meilleurs phonèmes, le troisième est considéré à sa place ;
— la catégorie phonétique de p'' ;
— le score acoustique de p et son rang par rapport aux scores de tous les modèles HMM.

Cette tâche de classification est basée sur les SVM (Support Vector Machines) qui ont été appliqués sur divers problématiques associées à la reconnaissance des formes (pattern recognition) (Vapnik, 1995; Scholkopf & Smola, 2001). Dans ce travail, la méthode de classification est appliquée sur deux classes : la discrimination entre les phonèmes normaux et anormaux (anomalies). Chaque phonème est caractérisé par l'ensemble de paramètres décrit auparavant. Tous les SVM utilisés ont des noyaux polynomiaux. Des modèles différents sont appris pour les hommes et les femmes ainsi que les différentes catégories phonétiques (consonnes sourdes, consonnes sonores, voyelles orales, voyelles nasales). Cette approche a été choisie afin d'affiner les spécificités de chaque catégorie phonétique. Les différents modèles sont appris en utilisant l'outil SVMlight (Joachims, 1999).

3 Corpus

Cette étude est basée sur deux corpus de parole. Le premier corpus (corpus Lysos.) contient 8 locuteurs dysarthriques et 6 contrôles. Les patients dysarthriques souffrent de maladies lysosomales résultant en une dysarthrie mixte et montrent plusieurs degrés de sévérité de dysarthrie (DSD). Tous les locuteurs ont lu le même texte, un conte pour enfant s'intitulant "Le cordonnier", le plus naturellement possible. Chacun a effectué entre 3 et 6 enregistrements sur des périodes séparées d'environ 6 mois. Tous les enregistrements des patients ont été annotés par un expert humain qui a identifié les anomalies au niveau du phonème. Grâce à cette annotation, ce corpus a été utilisé dans la phase d'apprentissage des modèles des phonèmes normaux et anormaux impliqués dans la phase de classification décrite dans la section 2.2. Le tableau 1 regroupe les informations liées à ce corpus.

TABLE 1 – Information liées au corpus Lysos. utilisé dans la phase d'apprentissage incluant le # de locuteurs, le # d'enregistrements et le degré de sévérité de dysarthrie (DSD) minimum et maximum par pathologie.

Pathologie	# de locuteurs	# d'enregistrements	(Min ;Max) DSD
Lysos.	8	35	(1.5 ;3.0)
Contrôles	6	17	-

Le second corpus, Typaloc, contient 28 locuteurs dysarthriques et 12 contrôles. Chaque locuteur a réalisé un enregistrement du même texte lu par le corpus Lysos. (le cordonnier) ainsi qu'un enregistrement additionnel de parole spontanée. Contrairement au premier corpus qui ne renferme que des patients atteints de maladies lysosomales, le corpus Typaloc présente des patients atteints de plusieurs pathologies et types de dysarthrie : la Sclérose Latérale Amyotrophique (SLA, dysarthrie mixte), la maladie de Parkinson (Park., dysarthrie hypokinétique) et l'Ataxie Cérébelleuse (AC, dysarthrie ataxique) réparties sur différents DSD. Un jury de 11 experts a évalué perceptivement tous les enregistrements des patients sur plusieurs critères de qualité de parole. Dans ce travail, nous nous concentrons sur le DSD, évalué sur une échelle de 0 à 3 (0 -pas de dysarthrie, 1 -dysarthrie légère,

TABLE 2 – Information sur le corpus Typaloc incluant le # de locuteurs, la moyenne de degré de sévérité de dysarthrie (DSD) et la moyenne ainsi que le (Min ;Max) des durées d'enregistrements (sec.) et # de phonèmes issus de l'alignement automatique par pathologie et style de parole.

		Parole lue			Parole spontanée		
Pathologie	# de locuteurs	Moy. DSD	Moy. dur. (Min ;Max)	Moy. # de phonèmes (Min ;Max)	Moy. DSD	Moy. dur. (Min ;Max)	Moy. # de phonèmes (Min ;Max)
SLA	12	2.0	111 (65 ;214)	532 (382 ;578)	2.0	102 (45 ;317)	463 (184 ;1089)
Park.	8	0.8	74 (48 ;122)	567 (550 ;599)	1.0	65 (40 ;109)	365 (315 ;404)
AC	8	1.3	107 (73 ;142)	569 (544 ;595)	1.2	78 (40 ;124)	423 (223 ;838)
Contrôles	12	-	70 (56 ;82)	558 (552 ;568)	-	522 (296 ;1028)	4072 (2762 ;7104)

2 -moyenne, 3 -sévère). Le tableau 2 résume les informations sur les locuteurs du corpus Typaloc groupées par pathologie et style de parole. Ces informations incluent le nombre de locuteurs, les valeurs moyennes de DSD, la moyenne, minimum et maximum de durée d'enregistrement (sec.) et le nombre de phonèmes issus de l'alignement automatique de la parole par locuteur. Il est intéressant de noter que les enregistrements de parole spontanée sont plus courts (durée et # de phonèmes) que ceux de la lecture indépendamment de la pathologie. Les patients sont en général peu enclins à parler spontanément à cause de leur pathologie évidemment et de l'effet de fatigue mais également à cause des techniques d'élicitation de parole qui ne sont pas toujours bien adaptées à ce type de locuteurs.

4 Résultats et discussions

Cette section décrit et discute le comportement de l'approche de détection automatique d'anomalies face aux paroles lue et spontanée issues du corpus Typaloc.
Comme rapporté dans (Laaridh *et al.*, 2015a) sur la parole lue, l'approche détecte plus d'anomalies chez les patients atteints de dysarthrie plus sévère. Ce comportement, bien que moins tranché, semble se conserver sur la parole spontanée. La figure 1 illustre les taux d'anomalies détectées par rapport aux DSD. En effet, la corrélation de Pearson entre ces taux et le DSD atteint 0.81 et 0.60 pour la parole lue et spontanée respectivement. Cette observation confirme que le système arrive à capturer l'évolution de la dysarthrie indépendamment de la tâche réalisée par le patient.
Le tableau 3 présente les taux d'anomalies automatiques sur les paroles lue et spontanée regroupés

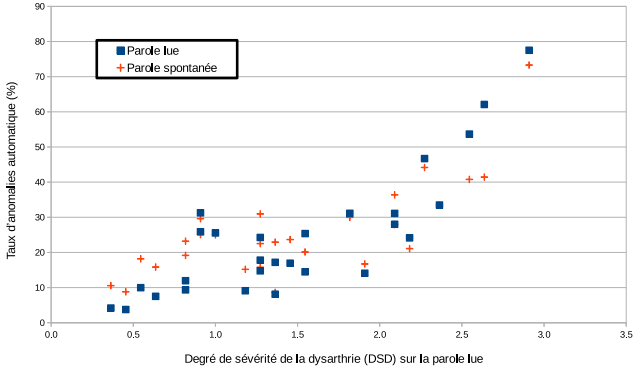


FIGURE 1 – Relation entre le taux d'anomalie automatique et le degré de sévérité de la dysarthrie (DSD) sur le corpus Typaloc

par population. On remarque que pour les contrôles, l'approche détecte plus d'anomalies sur la parole spontanée comparée à la parole lue. Cela peut être lié au fait que les modèles de phonèmes normaux et anormaux (section 2.2) ont été appris sur de la parole lue exclusivement. En effet, la parole spontanée peut présenter plus de variabilité acoustique qui est à la fois atypique (comparée à la parole lue) et non pathologique. Ces variations peuvent être liées au débit de parole plus rapide en spontanée, aux faux départs, hésitations ainsi qu'à des phénomènes de réduction plus fréquentes dans le cadre de parole spontanée. La figure 2 illustre la relation entre la différence des taux d'anomalies entre les deux styles de parole (taux d'anomalies sur la parole spontanée − taux d'anomalies sur la parole lue) et le DSD de la parole lue. Chaque point correspond à un locuteur (témoin ou patient). Observant le tableau 3 et la figure 2, nous remarquons que, comme pour les contrôles, l'approche détecte plus d'anomalies sur la parole spontanée pour les patients Park. et AC. Cette tendance est conforme aux résultats retrouvés dans (Van Lancker Sidtis *et al.*, 2012; Kempler & Van Lancker, 2002) sur des patients atteints de la maladie de Parkinson où la parole spontanée était moins intelligible et contenait plus de dysfluences que la parole lue. Par contre, les patients atteints de SLA présentent des taux d'anomalies similaires voire même inférieurs sur la parole spontanée par rapport à la lecture (32% et 36% respectivement). Dans notre corpus, ces patients souffrent des dysarthries les plus sévères (DSDs élevés). Une analyse plus approfondie de cette tendance est nécessaire pour vérifier si elle est liée aux caractéristiques intrinsèques de la SLA qui affecterait plus la tâche de lecture que celle de la parole spontanée. La deuxième hypothèse émise serait que ce phénomène est davantage lié au degré de sévérité élevé de la dysarthrie des patients SLA qu'à leur pathologie. En effet, la tâche de production de parole spontanée offre aux patients plus de "liberté" pour contrôler leur fatigue, leur débit de parole ainsi que les phonèmes et contextes à produire, ce qui pourrait donner moins d'anomalies au niveau phonème.

Finalement, le tableau 3 montre une augmentation plus importante des taux d'anomalies sur la parole spontanée chez les témoins que chez les patients dysarthriques (une augmentation relative de 154% pour les témoins contre 68%, 18% et -11% pour les Park., Cereb. et SLA respectivement) et la figure 2 suggère que la différence des taux d'anomalies entre les paroles spontanée et lue est inversement proportionnelle au DSD. Cela supposerait que les témoins changent davantage leurs productions selon le style de parole (ce qui résulte dans notre cas en plus d'anomalies détectées en spontanée vu la nature des modèles appris sur la lecture) alors que les patients (surtout les plus dysarthriques) perdent cette capacité à s'adapter aux différents styles et ont tendance à uniformiser leurs productions indépendamment de la tâche.

TABLE 3 – Taux d'anomalies moyen (%) par pathologie et style de parole calculés sur tous les phonèmes.

Pathologie	Parole lue	Parole spontanée
SLA	35.8	31.9
Park.	10.6	17.8
AC	20.6	24.4
Témoins	5.4	13.7

Le tableau 4 comporte les taux d'anomalies automatiques (%) par pathologie, catégorie phonétique et style de parole. Pour les contrôles, toutes les catégories phonétiques présentent des taux d'anomalies plus importants sur la parole spontanée comparée à la parole lue. Les fricatives enregistrent cependant une hausse nette des taux d'anomalies passant de 7% sur la parole lue à 23% sur la parole spontanée. Chez les Park., on trouve que les fricatives présentent les taux d'anomalies les plus importants pour les deux styles de parole. Il s'agirait donc davantage d'une amplification d'un comportement déjà observé (sur la parole lue) et lié à la dysarthrie parkinsonienne que de l'apparition d'un nouveau phénomène. Cependant, il est intéressant de constater que, contrairement aux autres catégories phonétiques où

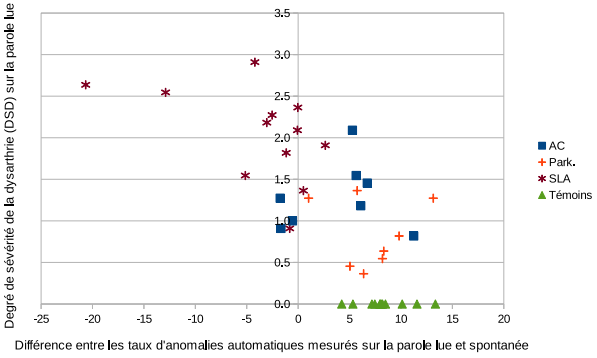


FIGURE 2 – Distribution des différences entre les taux d'anomalies sur les paroles spontanée et lue selon le DSD de la parole lue.

la hausse des taux d'anomalies sur la parole spontanée est légère, les fricatives présentent une augmentation absolue de 29% (112% relative). Considérant les patients atteints de SLA, bien que les taux d'anomalies soient plutôt stables entre les deux styles de parole, on remarque tout de même que les voyelles présentent moins d'anomalies en spontanée qu'en lecture (-4% et -18% pour les voyelles orales et nasales respectivement). La baisse du taux global d'anomalies observée chez les patients SLA entre la parole spontanée et la lecture est par conséquent le résultat de la baisse observée sur les voyelles, étant donné que les consonnes maintiennent des taux d'anomalies comparables entre les deux styles.

TABLE 4 – Taux d'anomalies moyen (%) calculé par pathologie, catégorie phonétique et style de parole (lue et spontanée).

	Parole lue				Parole spontanée			
Catégorie phonétique	Témoins	Park.	AC	SLA	Témoins	Park.	AC	SLA
Occlusives	7	12	22	37	11	16	24	35
Fricatives	7	26	48	37	23	55	53	38
Consonnes nasales	7	12	21	31	19	10	19	35
Consonne liquides	7	9	23	41	15	9	27	44
Voyelles Orales	2	6	10	33	11	10	13	29
Voyelles Nasales	4	8	20	44	8	13	13	26
Autres	10	16	26	46	19	13	26	37

Pour chaque population, une ANOVA à un facteur a été réalisée afin d'étudier l'effet du style de parole (2 niveaux : parole lue, parole spontanée). La figure 3 illustre les taux d'anomalies automatiques pour chaque population et style de parole. Chez les sujets contrôles ainsi que ceux atteints de Park., une différence significative est trouvée entre la parole lue et spontanée (($p<0.001$,$F(1,22)=28$) et ($p<0.05$, $F(1,14)=5.4$) respectivement). Ces différences sont encore plus visibles quand on se concentre seulement sur les fricatives (($p<0.001$,$F(1,22)=19$) et ($p<0.01$,$F(1,14)=14$) pour les locuteurs contrôles et Park. respectivement). La différence entre les deux styles de parole est moins flagrante pour les patients atteints de AC et SLA. Plus particulièrement pour les SLA, l'effet lié aux styles de parole peut être masqué par l'importante variabilité intra-pathologique observée sur cette population au niveau des taux d'anomalies automatique. De plus, cette variabilité est également observable au niveau des DSD des patients SLA présentant des dysarthrie légères, moyennes et sévères contrairement aux patients Park., atteints d'une dysarthrie légère.

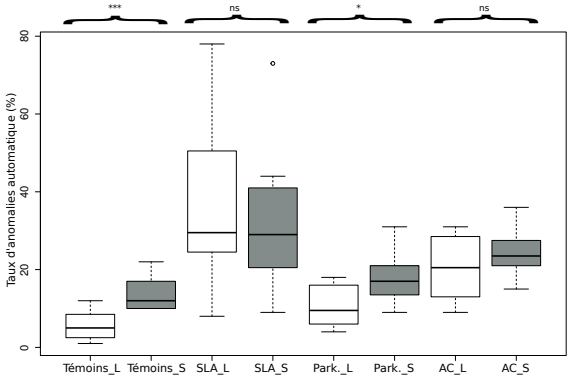


FIGURE 3 – Taux d'anomalies automatique par population et style de parole.

5 Conclusion

L'observation du comportement d'un système de détection automatique d'anomalies au niveau phonème sur la parole dysarthrique lue et spontanée a montré un effet de la tâche et du style de parole lié aux pathologies des patients. Les SLA, contrairement à toutes les autres populations (témoins, Park., AC), présentent plus d'anomalies sur la parole lue que la parole spontanée. Globalement, ce sont les témoins qui montrent le plus de changement selon le style de parole, et chez les patients, plus la dysarthrie est sévère moins il y a de différences entre les deux styles. Une hypothèse résultante pourrait alors être que les témoins adaptent leur production selon le style de parole alors que les patients dysarthriques perdent graduellement cette capacité à s'adapter aux différents styles de parole. Chez les contrôles et les Park., les fricatives montrent une importante hausse du taux d'anomalies dans le cadre de la parole spontanée.
De prochains travaux examineront l'effet de la localisation des phonèmes (première, deuxième...etc syllabe) sur le processus de détection d'anomalies.

Références

BROWN A. & DOCHERTY G. J. (1995). Phonetic variation in dysarthric speech as a function of sampling task. *International Journal of Language & Communication Disorders*, **30**(1), 17–35.

CARMICHAEL J. (2007). *Introducing objective acoustic metrics for the Frenchay Dysarthria Assessment procedure*. Ph.d. dissertation, university of sheffield.

CHANDOLA V., BANERJEE A. & KUMAR V. (2007). *Anomaly detection : a survey*. University of Minnesota (USA).

CHRISTENSEN H., CASANUEVA I., CUNNINGHAM S., GREEN P. & HAIN T. (2013). homeservice : Voice-enabled assistive technology in the home using cloud-based automatic speech recognition. In *4th Workshop on Speech and Language Processing for Assistive Technologies*, p. 29–34.

DARLEY F. L., ARONSON A. E. & BROWN J. R. (1969). Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research*, **12**, 246–269.

DARLEY F. L., ARONSON A. E. & BROWN J. R. (1975). *Motor speech disorders*. Philadelphia : W. B. Saunders and Co.

DUFFY J. R. (2005). *Motor speech disorders : substrates, differential diagnosis and management*. Motsby- Yearbook, St Louis, 2nd edition.

ENDERBY P. (1983). Frenchay dysarthric assessment. *Pro-Ed, Texas*.

FREDOUILLE C. & POUCHOULIN G. (2011). Automatic detection of abnormal zones in pathological speech. In *Intl Congress of Phonetic Sciences (ICPHS'11)*, Hong Kong.

GALLIANO S., GEOFFROIS E., MOSTEFA D., CHOUKRI K., BONASTRE J.-F. & GRAVIER G. (2005). ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In *Proceedings of Interspeech'05*, p. 1149–1152.

GREEN J. R., YUNUSOVA Y., KURUVILLA M. S., WANG J., PATTEE G. L., SYNHORSTI L., ZINMAN L. & BERRY J. D. (2013). Bulbar and speech motor assessment in ALS : Challenges and future directions. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, **14(7–8)**, 494–500.

JOACHIMS T. (1999). Making large-scale SVM learning practical. In B. SCHÖLKOPF, C. BURGES & A. SMOLA, Eds., *Advances in Kernel Methods - Support Vector Learning*, chapter 11, p. 169–184. Cambridge, MA : MIT Press.

KEMPLER D. & VAN LANCKER D. (2002). Effect of speech task on intelligibility in dysarthria : a case study of Parkinson's disease. *Brain and language*, **80**(3), 449–464.

KENT R. D., WEISMER G., KENT J. F., VORPERIAN H. K. & DUFFY J. R. (1999). Acoustic studies of dysarthric speech : Methods, progress, and potential. *The Journal of Communication Disorders*, **32 :3**, 141–186.

KIM M. & KIM H. (2012). Automatic assessment of dysarthric speech intelligibility based on selected phonetic quality features. In *Computers Helping People with Special Needs*, volume 7383 of *Lecture Notes in Computer Science*, p. 447–450.

LAARIDH I., FREDOUILLE C. & MEUNIER C. (2015a). Automatic detection of phone-based anomalies in dysarthric speech. *ACM Transactions on accessible computing*, **6**(3), 9 :1–9 :24.

LAARIDH I., FREDOUILLE C. & MEUNIER C. (2015b). Automatic speech processing for dysarthria : A study of inter-pathology variability. In *Intl Congress of Phonetic Sciences (ICPHS'15)*, Glasgow.

LOWIT A. & KENT R. D. (2010). *Assessment of motor speech disorders*, volume 1. Plural publishing.

MIDDAG C., MARTENS J.-P., NUFFELEN G. V. & BODT M. D. (2009). Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on Applied Signal Processing*, **2009(1)**.

PARKER M., CUNNINGHAM S., ENDERBY P., HAWLEY M. & GREEN P. (2006). Automatic speech recognition and training for severely dysarthric users of assistive technology : the stardust project. *Clinical Linguistics and Phonetics*, **20(2–3)**, 149–156.

SCHOLKOPF B. & SMOLA A. J. (2001). *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA : MIT Press.

TOMIK B. & GUILOFF J. (2010). Dysarthria in amyotrophic lateral sclerosis : a review. *Amyotrophic Lateral Sclerosis*, **11 (1–2)**, 4–15.

VAN LANCKER SIDTIS D., CAMERON K. & SIDTIS J. J. (2012). Dramatic effects of speech task on motor and linguistic planning in severely dysfluent parkinsonian speech. *Clinical linguistics & phonetics*, **26**(8), 695–711.

VAPNIK V. (1995). *The Nature of Statistical Learning Theory*. New York, NY, USA : Springer-Verlag New York, Inc.

YORKSTON K. M., STRAND E. & KENNEDY M. (1996). Comprehensibility of dysarthric speech : implications for assessment and treatment planning. *American Journal of Speech Language Pathology*, **55**, 55–66.

Effet de l'input auditif sur la production de voyelles orales : étude acoustique chez des enfants normo-entendants et des enfants porteurs d'implants cochléaires âgés de 5 à 11 ans

Bénédicte Grandon, Anne Vilain
Gipsa-Lab, Université Grenoble Alpes, 1180 Avenue Centrale, 38040 Grenoble Cedex 9, France
Benedicte.Grandon@gipsa-lab.grenoble-inp.fr, Anne.Vilain@gipsa-lab.grenoble-inp.fr

RESUME

Treize enfants porteurs d'implants cochléaires (CI) et vingt enfants normo-entendants (NH) ont été enregistrés dans deux conditions : répétition de mots avec un modèle audio et production des mêmes mots sans modèle audio. Notre but était d'étudier l'effet de l'input audio sur la hauteur, l'antériorité et la dispersion des dix voyelles orales du français chez ces deux populations d'enfants. Les résultats de notre étude acoustique indiquent que : (1) l'input immédiat n'influence que la hauteur du /a/ chez les enfants NH, (2) les enfants CI produisent des voyelles /y/, /ø/, /œ/ plus postérieures que les enfants NH mais que cette différence diminue à mesure que la durée d'utilisation de l'implant augmente, et (3) la dispersion de /y/, /ø/, /œ/ est plus grande chez les enfants CI que chez les enfants NH.

ABSTRACT

Effect of audio input on vowel production: an acoustic study in 5- to 11-year old normal-hearing and cochlear implanted children

Thirteen cochlear implanted children (CI) and twenty normal-hearing children (NH) were recorded in two conditions: repetition of words with an audio model, and production of the same words without audio model. Our goal was to study the effect of audio input on the height, anteriority and dispersion of the 10 French oral vowels in these two populations of children. The results of our acoustic study show that (1) immediate input only affects height in /a/ in NH children, that (2) CI children have a more posterior realization of /y/, /ø/, /œ/ than NH children, with this difference decreasing with a longer use of the cochlear implant and that (3) the dispersion of /y/, /ø/, /œ/ is bigger in CI children compared to NH children.

MOTS-CLES : voyelles, production, enfants, implant cochléaire.
KEYWORDS: vowels, production, children, cochlear implant.

1 Introduction

L'enjeu principal de l'implant cochléaire est de permettre à son utilisateur de percevoir des sons, en particulier des sons de parole, qu'il ne percevrait pas par ailleurs, pour ensuite pouvoir communiquer de façon intelligible. L'implant cochléaire capte et décompose les sons puis en transmet une reconstitution plus ou moins partielle vers le nerf auditif. L'implant cochléaire

pédiatrique permet ainsi à des enfants sourds pré-linguaux d'acquérir une représentation phonologique des sons de parole à partir des fréquences de sons codées par l'implant. Dans cette étude, nous nous intéressons à la production de voyelles par des enfants sourds pré-linguaux, porteurs d'implants cochléaires, âgés de 6 à 11 ans plusieurs années après implantation. Notre but est de comprendre quelles sont les difficultés persistantes rencontrées par ces enfants en comparant leurs productions à celles d'enfants normo-entendants du même âge.

La plupart des études acoustiques récentes sur la production de voyelles par des enfants implantés s'intéressent à l'évolution avant et dans la première année qui suit l'implantation (par exemple Ertmer, 2001, Hocevar-Bolthezar et al., 2008), avec implant activé ou désactivé (Poissant et al., 2006) ou en comparant leurs productions avec des groupes contrôles d'enfants normo-entendants (Horga, Liker, 2006, Baudonck et al., 2011). Ces études utilisent des techniques différentes d'élicitation de parole : répétition de mots ou de syllabes, lecture, parole spontanée, production de mots sans modèle audio… Ces études sur les productions de voyelles chez des enfants CI font le choix d'une méthode mais aucune ne compare des productions obtenues avec plusieurs méthodes. Les résultats de ces études sont assez disparates, pour les caractéristiques acoustiques étudiées (F1, F2 et taille de l'espace vocalique) : (1) F1 de /u/ plus variable chez les enfants porteurs d'implants que chez les enfants normo-entendants (Baudonck et al., 2011), distinction claire entre les F1 de /i-u/ par rapport à ceux de /ɑ-æ/ mais étendue plus grande entre /ɑ/ et /æ/ qu'entre /i/ et /u/ (Ertmer, 2001), diminution des F1 de /u/ et /i/ après l'implantation (Hocevar-Bolthezar et al., 2008), (2) F2 de /a/ plus bas chez les enfants porteurs d'implants que chez les enfants normo-entendants, F2 de /i/ et /u/ similaires pour les deux groupes (Baudonck et al., 2011), étendue de F2 similaire pour /i/, /u/ et /æ/ et étendue de F2 de /a/ plus petite que celle de F2 de /i/, /u/ et /æ/ chez l'enfant porteur d'implant de l'étude (Ertmer, 2001), et (3) espace vocalique plus grand chez les enfants porteurs d'implants que chez les enfants normo-entendants (Baudonck et al., 2011) ou de taille comparable chez les deux groupes (Horga, Liker, 2006). Ces résultats disparates peuvent être expliqués à la fois par des méthodologies de collecte et d'analyses des données différentes (avec ou sans normalisation des formants) et par des systèmes vocaliques différents selon les langues (Anglais américain pour Ertmer, 2001 et Poissant et al., 2006, flamand pour Baudonck et al., 2011, slovène pour Hocevar-Bolthezar et al., 2008 et croate pour Horga, Liker, 2001).

A travers une étude acoustique de la production de voyelles, nous cherchons à comprendre quels sont les effets à long-terme de l'utilisation de l'implant cochléaire sur la production de voyelles, en nous intéressant en particulier aux capacités des enfants à utiliser l'input immédiat qu'ils reçoivent pour percevoir et reproduire des voyelles. La perception des voyelles est liée principalement à des caractéristiques acoustiques fréquentielles (formants), et l'implant cochléaire ne reproduit que partiellement l'étendue des fréquences des sons de parole. Nous pouvons donc nous attendre à une production de parole plus variable chez les enfants porteurs d'implants que chez des enfants normo-entendants de même âge, ainsi qu'à une production plus variable lorsque les enfants n'ont pas de modèle audio immédiatement à disposition. Plusieurs études sur la production de voyelles par des adultes porteurs d'implants (étude en allemand de Neumeyer et al., 2010) ou de consonnes par des enfants sourds utilisateurs d'aides auditives traditionnelles (étude en anglais américain de Geffner, 1980) ou d'implants cochléaires (étude en français québécois de Gaul-Bouchard et al., 2007) ont mis en évidence l'effet du degré de visibilité des phonèmes sur la production : des phonèmes (consonnes ou voyelles) dont la production est jugée ambiguë visuellement sont produits avec plus de difficultés par des personnes sourdes utilisant aides auditives traditionnelles ou implants cochléaires. Nous nous attendons donc à obtenir (1) des différences faibles de F1 entre les deux groupes, puisque la hauteur de la voyelle n'est pas une caractéristique visuellement ambiguë, (2) des différences marquées de F2 entre les deux groupes, qui seraient liées à l'ambiguïté visuelle entre

voyelles antérieures arrondies /y/, /ø/, /œ/, et leurs correspondantes postérieures /u/, /o/ /ɔ/, (3) une dispersion plus grande de chaque voyelle chez les enfants porteurs d'implants et (4) une variabilité plus grande pour les deux groupes lorsque la production se fait sans modèle audio immédiatement disponible mais à partir d'une représentation phonologique de la voyelle, cette variabilité pouvant être corrélée à l'âge à l'implantation ou à la durée d'utilisation de l'implant chez les enfants porteurs d'implants cochléaires.

2 Méthode

Pour cette étude, nous avons enregistré les productions de parole de deux groupes d'enfants (normo-entendants (NH) et porteurs d'implants (CI)), dans une tâche de répétition et une tâche de production. Notre but est de comprendre (1) quel est l'impact de l'implant (2) quel est l'effet de l'input immédiat sur la production de parole et (3) quels facteurs influencent la production de parole lorsque les enfants ont une perception partielle des sons.

2.1 Participants

Les participants étaient 20 enfants normo-entendants (NH, âge moyen : 7;8 ans (5;7-10;6 ans)) et 13 enfants sourds pré- ou périlinguistiques porteurs d'implants cochléaires (CI, âge moyen : 8;2 ans (6;6 -10;7 ans), âge moyen du diagnostic de la surdité : 1;6 ans (0;7 ans – 3 ans), âge à l'implantation moyen 3;1 ans (1;6-6;6 ans), durée moyenne d'utilisation de l'implant : 5;2 ans (2;2-9;1 ans)) : 7 enfants étaient porteurs d'un implant et 6 enfants étaient porteurs de deux implants. Les deux groupes d'enfants étaient appariés en âge (t-test : p=0.3117). Tous les enfants normo-entendants (20) étaient originaires d'Isère, et les enfants porteurs d'implants étaient originaires d'Isère (5), Rhône (5), Saône-et-Loire (2) et Haute-Savoie (1).

2.2 Tâches

Les enfants ont participé à deux tâches différentes : pour la tâche de répétition, un modèle audio des mots du corpus, produits par une locutrice francophone enregistrée préalablement était présenté avec une image et pour la tâche de production les mêmes images étaient présentées sans modèle audio. Chaque enfant a effectué chaque tâche deux fois, à l'exception d'un enfant CI qui a effectué chaque tâche une seule fois. Les mots étaient présentés en ordre aléatoire, l'ordre des tâches était le même (répétition, puis production) pour tous les enfants, la tâche de répétition servant également d'entrainement à la tâche de production.

2.3 Corpus

Le corpus était constitué d'une liste de 10 mots mono- et bisyllabiques du français (objets, animaux, etc.) connus d'enfants de 5 à 10 ans, comprenant les 10 voyelles orales du français /i, e, ɛ, y, ø, œ, a, u, o, ɔ/ dans une séquence CV initiale de mot. Le contexte des voyelles-cibles était soit /#b_/ soit /#b_C/ (ex : « bateau », « bœuf »). Les mots enregistrés pour être utilisés comme modèles audio étaient accentués sur la première syllabe, de façon à éliminer dans la mesure du possible une variation de production liée à des différences d'accentuation.

2.4 Enregistrements

Les enregistrements des enfants NH ont tous été réalisés en chambre sourde, au laboratoire Gipsa-Lab. Les enregistrements des enfants CI ont été réalisés au CHU de Grenoble (2 enfants) et à l'hôpital E. Herriot de Lyon (11 enfants). Les enregistrements ont été réalisés avec un enregistreur numérique de type Marantz PMD 670 (mono, fréquence d'échantillonnage 44100 Hz, 16 bits) et un microphone externe, placé à environ 40 cm des enfants.

2.5 Analyses des données

– Traitement des enregistrements

Les enregistrements ont été tout d'abord isolés en mots, puis segmentés et annotés sous Praat (Boersma, Weenink, 2003) : chaque voyelle-cible a été annotée, en utilisant une notation SAMPA (correspondance API-SAMPA: /i, e, ɛ, y, ø, œ, a, u, o, ɔ/ > /i, e, E, y, 2, 9, a, u, o, O/). Les productions inexploitables de chaque mot et chaque voyelle ont également été identifiées (mot non-produit ou substitué par un autre mot, dévoisement ou élision de voyelle…), pour pouvoir être exclues des analyses acoustiques. Exemple de production inexploitable : substitution de « bœuf » /bœf/ par « vache » /vaʃ/.

– F1 et F2 : mesures et normalisation

A partir des annotations réalisées, nous avons mesuré le premier et le deuxième formant de chaque voyelle en un point, au centre de la voyelle. Ces mesures ont été réalisées automatiquement, à l'aide d'un script Praat (Boersma, Weenink, 2003), puis vérifiées manuellement, pour s'assurer de leur exactitude.

Les valeurs de F1 et F2 ainsi obtenues ont été normalisées selon la méthode des scores-z de Lobanov (Lobanov, 1971). Notre étude a pour but de comparer la production de voyelles chez deux groupes d'enfants ayant des capacités perceptives différentes, et nous avons donc choisi une méthode de normalisation qui nous permet d'éliminer toute variance liée à l'âge des enfants (résultant d'une différence de taille de conduit vocal), en conservant la variance liée aux capacités perceptives (Grandon et al., 2014). La méthode de Lobanov (Lobanov, 1971) consiste à calculer pour chaque formant une distance par rapport à la moyenne de ce formant pour toutes les voyelles de l'espace vocalique de ce locuteur, puis à diviser cette distance par l'écart-type de ce formant.

– Dispersion de chaque catégorie vocalique

Nous avons calculé, à partir des mesures normalisées de F1 et F2 de chaque occurrence, la distance euclidienne en deux dimensions (dans l'espace F1-F2) de cette occurrence par rapport au centre de la catégorie correspondante, pour chaque groupe et chaque condition (Répétition et Production), avec la formule suivante (avec x : occurrence de la voyelle, gr : groupe, cond : condition et i : catégorie vocalique) :

$$dist(x)_i = \sqrt{(F1(x)_i - F1(moy)_{gr,cond,i})^2 + (F2(x)_i - F1(moy)_{gr,cond,i})^2}$$

Cette distance euclidienne nous permet de comparer les dispersions de chaque catégorie vocalique pour chaque groupe et chaque condition. Toutes les analyses de la dispersion ci-après utilisent ces calculs de la distance euclidienne de chaque occurrence produite par rapport au centre de la catégorie vocalique correspondante.

2.6 Analyses statistiques

Tous les graphiques et analyses statistiques ont été réalisés avec le logiciel R (R Development Core Team, 2012). Nous avons dans un premier temps utilisé des modèles linéaires (fonctions *stepAIC* et *lme* sur R), pour étudier l'effet de plusieurs facteurs, seuls ou en interaction (hauteur/antériorité de la voyelle, condition, groupe, âge chronologique, âge à l'implantation, âge auditif) sur les variables à expliquer (F1, F2, dispersion F1-F2 de chaque catégorie vocalique). L'âge auditif correspond à la durée d'utilisation de l'implant par les enfants, c'est à dire la durée entre l'implantation et le moment où ils ont participé à notre étude. Nous avons regroupé les voyelles en quatre degrés de hauteur pour l'étude du F1 : voyelles hautes /i/, /y/, /u/ (/i/, /y/, /u/ en SAMPA sur les figures), voyelles mi-hautes /e/, /ø/, /o/ (/e/, /2/, /o/ en SAMPA), voyelles mi-basses /ɛ/, /œ/, /ɔ/ (/E/, /9/, /O/ en SAMPA), voyelle basse /a/ (/a/ en SAMPA), et en quatre degrés d'antériorité pour l'étude du F2 : voyelles antérieures non-arrondies /i/, /e/, /ɛ/ (/i/, /e/, /E/ en SAMPA), voyelles antérieures arrondies /y/, /ø/, /œ/ (/y/, /2/, /9/ en SAMPA), voyelle centrale /a/ (/a/ en SAMPA) et voyelles postérieures /u/, /o/, /ɔ/ (/u/, /o/, /O/ en SAMPA).

Ensuite, nous avons comparé, pour chaque voyelle, les moyennes des variables (F1, F2, dispersion F1-F2 de chaque voyelle) en fonction du groupe (CI vs NH), puis de la condition (production vs répétition). Pour ces comparaisons multiples, nous utilisons les fonctions *lsmeans* et *multcomp* sur R.

3 Résultats

3.1 F1

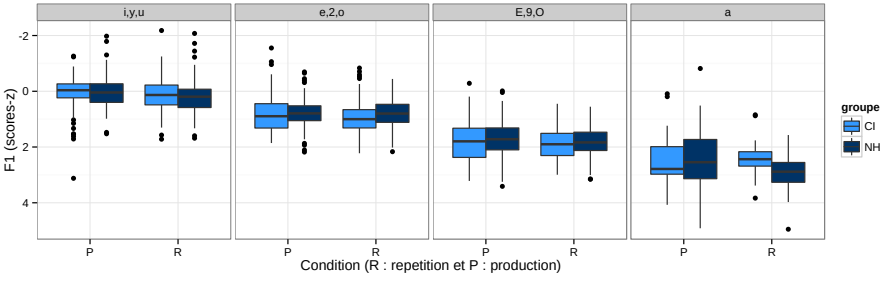


FIGURE 1: valeurs de F1 par groupe (enfants CI et NH), par condition (production et répétition), et par degré de hauteur de la voyelle

– Effets de la condition et du groupe sur la réalisation de la hauteur des voyelles

Les tests statistiques (modèle linéaire à effets mixtes, avec un effet aléatoire du sujet) indiquent un effet de la hauteur ($p<.0001$), de la condition ($p<.0001$), ainsi qu'un effet de l'interaction groupe:hauteur sur la réalisation du F1 ($p=0.0473$). Les deux groupes d'enfants réalisent la hauteur différemment (les voyelles /i/, /y/ et /u/ sont plus hautes chez les enfants CI, les voyelles /e/, /ø/, /o/ et /i/, /e/, /ɛ/ sont plus basses chez les enfants CI que chez les enfants NH dans les deux conditions, le /a/ des enfants CI est plus bas que celui des enfants NH en production et plus haut en répétition) mais le test de comparaisons multiples indique que cette différence de hauteur entre les deux groupes n'est significative que pour la voyelle /a/, en répétition ($p=.0409$). De même, la condition (répétition vs production) n'influence que la production de la voyelle /a/, chez les enfants NH ($p=.0012$).

– Facteurs influençant la réalisation de la hauteur des voyelles

Les tests statistiques indiquent qu'il n'y a pas d'effet de l'âge chronologique sur le F1 des enfants NH (p>.05), ce qui était attendu en raison du choix d'une méthode de normalisation qui élimine la variance liée à l'âge chronologique.

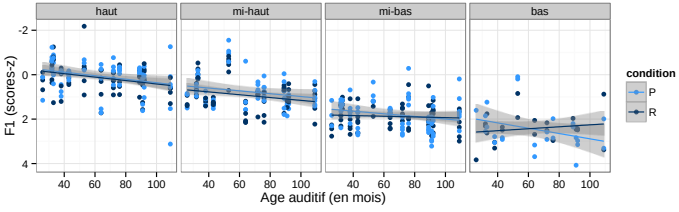


FIGURE 2 : F1 en fonction de l'âge auditif par condition et par degré de hauteur (enfants CI)

Chez les enfants CI, il n'y a pas d'effet global de l'âge chronologique sur le F1 (p>.05), mais un effet de l'interaction voyelle : âge auditif (p=.0064) et de l'interaction voyelle : âge à l'implantation (p =.0295) sur le F1. Ces effets ne sont pas les mêmes pour tous les degrés de hauteur. Pour les voyelles hautes /i/, /y/ et /u/, il y a un effet de l'âge d'audition sur le F1 (p=.0325), qui correspond à une augmentation du F1 et donc un abaissement de la hauteur des voyelles hautes. Pour les voyelles mi-hautes /e/, /ø/, /o/, nous observons un effet de la condition (p=.0147) ainsi que de l'interaction condition : âge à l'implantation (p=.0034) sur le F1, qui correspond à un abaissement des voyelles mi-hautes, qui est plus marqué en répétition. Pour les voyelles mi-basses /ɛ/, /œ/, /ɔ/, nous n'observons pas d'effet de l'âge à l'implantation ni de l'âge auditif sur le F1. Pour la voyelle /a/, nous observons un effet de l'interaction condition : âge auditif (p=.0218) sur le F1, ce qui correspond à un abaissement de la voyelle /a/ lors de la tâche de répétition.

3.2 F2

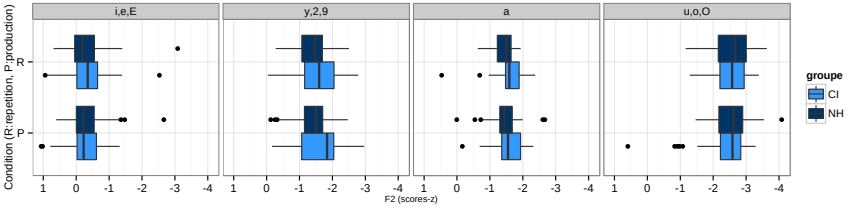


FIGURE 3 : valeurs de F2 par groupe (enfants CI et NH), par condition (production et répétition), et par degré d'antériorité de la voyelle

– Effets de la condition et du groupe sur la réalisation de l'antériorité des voyelles

Les tests statistiques (modèle linéaire à effets mixtes, avec un effet aléatoire du sujet) indiquent un effet de l'antériorité (p<.0001) et de l'interaction groupe:antériorité (p=.0005) sur la réalisation du F2 : la figure 3 indique que les enfants CI produisent les voyelles antérieures arrondies /y/, /ø/, /œ/ et la voyelle /a/ de façon plus postérieure que les enfants NH. Les tests de comparaisons multiples montrent que cette différence constatée graphiquement n'est pas significative pour /a/ (p=.38468) mais elle l'est pour le groupe des voyelles antérieures arrondies /y/, /ø/, /œ/ (p=.00908). A la différence des résultats sur le F1, la condition n'a pas d'effet sur la réalisation du F2.

– Facteurs influençant la réalisation de l'antériorité des voyelles

Comme pour le F1, il n'y a pas d'effet de l'âge chronologique sur le F2 des enfants NH (p>.05), en raison de la méthode de normalisation choisie.

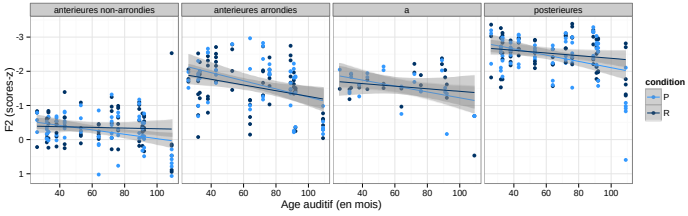


FIGURE 4 : F2 en fonction de l'âge auditif par condition et par degré d'antériorité (enfants CI)

Les tests statistiques ne montrent pas d'effet de l'âge chronologique, de l'âge auditif ni de l'âge à l'implantation sur le F2 des voyelles antérieures non-arrondies, les voyelles postérieures et la voyelle /a/. Cependant, nous observons un effet de l'âge auditif sur le F2 des voyelles antérieures arrondies (p=.0343) : le F2 augmente avec l'âge auditif, ce qui signifie que plus les enfants CI ont une utilisation longue de l'implant, plus le F2 de leurs voyelles /y/, /ø/, /œ/ se rapproche du F2 des mêmes voyelles produites par les enfants NH.

3.3 Dispersion de chaque catégorie vocalique

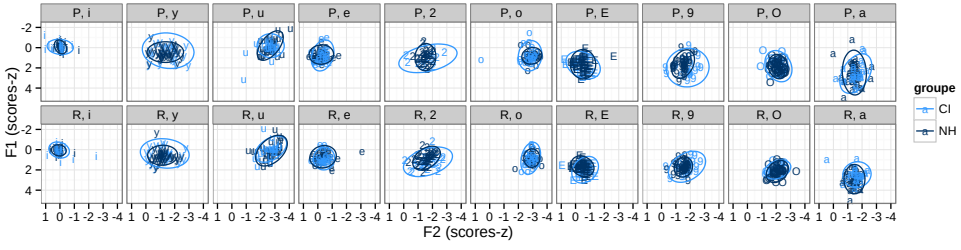


FIGURE 5 : représentation F1-F2 de chaque catégorie vocalique, par groupe (enfants CI et NH), par condition (production et répétition), ellipses à 95% de confiance.

Nous observons, sur la figure 5 des distributions spatiales différentes de certaines catégories vocaliques, selon les groupes et les conditions. Il semble que les catégories soient réalisées avec plus de variabilité lorsque les enfants n'avaient pas de modèle audio (tâche de production). Par ailleurs les voyelles produites par les enfants CI semblent avoir une plus grande dispersion que celles des enfants NH. Ces différences sont plus marquées pour certaines voyelles.

Les tests statistiques que nous avons effectués sur la dispersion des catégories vocaliques confirment certaines de ces observations : il y a une différence significative entre les enfants CI et NH pour les voyelles antérieures arrondies (p<.0001) : ces voyelles produites par les enfants CI ont une plus grande dispersion dans l'espace. Les productions des autres voyelles (antérieures non-arrondies, postérieures et voyelles /a/) ne sont pas significativement différentes entre les deux groupes., l'observation graphique de légères différences de dispersion entre les deux conditions ne se confirme pas statistiquement, et ce, pour toutes les voyelles.

– Facteurs influençant la dispersion des catégories vocaliques

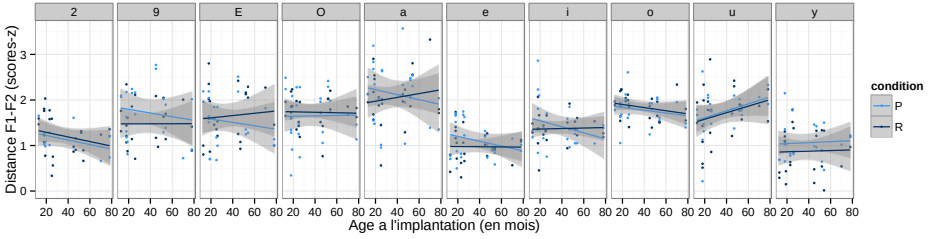


FIGURE 6 : dispersion des catégories vocaliques, en fonction de l'âge à l'implantation par condition et par catégorie (enfants CI)

Les tests statistiques montrent qu'il n'y a pas d'effet de l'âge chronologique sur la dispersion des catégories vocaliques. Seule la voyelle /u/ est affectée par une interaction condition : âge chronologique (p=.018) : la dispersion de la catégorie augmente légèrement chez les enfants NH mais diminue fortement chez les enfants CI en fonction de l'âge chronologique lors de la tâche de répétition. Par ailleurs, chez les CI, l'âge auditif n'a pas d'effet sur la dispersion des catégories vocaliques et seule la dispersion des voyelles /y/ et /o/ est affectée par l'âge à l'implantation : pour /y/, l'interaction condition : âge à l'implantation a un effet significatif sur la dispersion (p=.0439) et pour /o/, l'âge à l'implantation a un effet significatif sur la dispersion de la catégorie (p=.0320) : plus l'implantation est tardive, moins la dispersion de la voyelle /o/ est grande, alors que la voyelle /y/ est nettement plus dispersée en production qu'en répétition, et légèrement plus dispersée lorsque l'implantation est tardive.

4 Discussion/Conclusion

A travers cette étude acoustique, nous avons mis en évidence que (1) l'implantation permet aux enfants CI de percevoir certaines caractéristiques acoustiques des voyelles qui leur permettent ensuite de produire des voyelles globalement comparables en terme de hauteur, d'antériorité et de dispersion à celles d'enfants normo-entendants du même âge, sauf pour ce qui concerne la série des voyelles antérieures arrondies : les enfants CI ont tendance à produire ces voyelles de façon plus postérieure et avec plus de variabilité que les enfants NH. Une hypothèse explicative serait que ces voyelles sont mal distinguées phonologiquement chez les enfants CI des voyelles postérieures arrondies, en raison des similitudes en termes d'informations visuelles (2) l'input immédiat n'a que peu d'influence sur les productions de voyelles par les deux groupes, la voyelle /a/ étant la seule voyelle pour laquelle une différence de hauteur entre les deux conditions (production et répétition) est observée et statistiquement significative. Ceci semble indiquer que les patrons de production des catégories vocaliques sont bien stabilisés, autant chez les CI que les NH, dès l'âge de 5 ans, et (3) les différences observées au niveau de l'antériorité des voyelles antérieures arrondies ne semblent pas être dues à l'âge d'implantation mais à la durée d'utilisation de l'implant : la tendance des enfants CI à produire les voyelles antérieures arrondies de façon plus postérieure que les enfants NH tend à diminuer lorsque la durée d'utilisation de l'implant augmente.

Remerciements

Ce projet est financé par une Allocation Doctorale de Recherche de la Région Rhône-Alpes (projet ARC2, porté par A. Vilain et H. Loevenbruck). Nous remercions les Professeurs S. Schmerber (CHU Grenoble) et E. Truy (Hopital E. Herriot de Lyon), ainsi que les orthophonistes du CHU de

Grenoble et la coordinatrice du centre d'implantation de l'Hôpital E. Herriot de Lyon. Nous remercions également tous les enfants et leurs parents.

Références

Baudonck N., Van Lierde K., Dhooge I., Corthals P. (2011). A comparison of vowel productions in prelingually deaf children using cochlear implants, severe hearing-impaired children using conventional hearing aids and normal-hearing children. *Folia Phoniatrica et Logopaedica*, 63(3), 154-160.

Boersma P., Weenink D. (2013). Praat: doing phonetics by computer (Version 5.3.80)

Ertmer D. J. (2001). Emergence of a vowel system in a young cochlear implant recipient. *Journal of Speech, Language, and Hearing Research, 44(4),* 803-813.

Gaul-Bouchard M. E., Le Normand M. T., Cohen H. (2007). Production of consonants by prelinguistically deaf children with cochlear implants. *Clinical linguistics & phonetics*, *21(11-12*), 875-884.

Geffner D. (1980). Feature characteristics of spontaneous speech production in young deaf children. *Journal of Communication Disorders, 13(6)*, 443-454.

Grandon B., Vilain A., Lœvenbruck H., Schmerber S. (2014). Vowel spaces in French children wearing cochlear implants. Actes de : *LSCD 2014 - Workshop on Late Stages in Speech and Communication Development*, 93-95.

Hocevar-Boltezar I., Boltezar M., Zargi M. (2008). The influence of cochlear implantation on vowel articulation. *Wiener klinische Wochenschrift, 120(7-8),* 228-233.

Horga D., Liker M. (2006). Voice and pronunciation of cochlear implant speakers. *Clinical linguistics & phonetics*, *20(2-3),* 211-217.

Lobanov B. M. (1971). Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America, 49(2B),* 606-608.

Neumeyer V., Harrington, J., Draxler, C. (2010). An acoustic analysis of the vowel space in young and old cochlear-implant speakers. *Clinical linguistics & phonetics*, *24(9),* 734-741.

Poissant S. F., Peters K. A., Robb M. P. (2006). Acoustic and perceptual appraisal of speech production in pediatric cochlear implant users. *International journal of pediatric otorhinolaryngology*, *70(7),* 1195-1203.

R Development Core Team (2012). R: A language and environment for statistical computing. R, Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.Rproject.org/.

Effet de la fréquence d'usage sur l'élision du schwa des clitiques : étude d'un corpus d'interactions naturelles

Loïc Liégeois[1, 2]
(1) CLILLAC-ARP, 5 rue Thomas Mann, 75013 Paris, France
(2) LLF, 5 rue Thomas Mann, 75013 Paris, France
loic.liegeois@univ-paris-diderot.fr

RESUME

Cette étude s'intéresse à l'influence d'un facteur d'usage, à savoir la fréquence des formes, sur la (non) production des schwas des clitiques. Dans cet objectif, nous nous appuyons sur un corpus d'interactions entre adultes recueillies en situation naturelle : les enregistrements, réalisés au domicile de nos six sujets, ont été récoltés au cours de scènes de vie quotidienne. Les données présentées au cours de nos analyses corroborent les résultats exposés dans de précédents travaux au sujet des schwas initiaux de polysyllabes. En effet, il s'avère que la fréquence d'emploi des collocations "Clitique + X" a un effet significatif sur les taux d'élision relevés dans les productions de nos sujets.

ABSTRACT

Frequency effect on schwa elision in clitics: a corpus based study
This study deals with the effect of a usage factor, token frequency, on the elision of schwa in clitics. For this purpose, we base our study on a corpus of interactions between adults collected in natural settings: the recordings, realised at the speakers' home, were collected during scenes of everyday life. Our results corroborate those of previous studies concerning initial schwa in polysyllabic words. Indeed, the frequency of "Clitic + X" collocations has a significant effect on elision rates observed in the productions of the six speakers recorded.

MOTS-CLES : schwa, clitiques, effets de fréquence, corpus d'interactions naturelles
KEYWORDS: schwa, clitics, frequency effects, ecological corpus

1 Introduction

Le schwa, considéré comme une voyelle particulière du français, a suscité un nombre relativement important de travaux portant essentiellement sur ses particularités phonétiques et son statut phonologique. En ce qui concerne les facteurs de (non) production du schwa, la multidimensionnalité du phénomène a amené les chercheurs à se pencher sur une grande variété de conditions : phonétiques et phonologiques, bien sûr, mais également syntaxiques, discursives ou sociolinguistiques. Notre étude propose de se focaliser sur un facteur en lien avec l'usage, la fréquence d'usage des formes. Dans cet objectif, nous nous appuierons sur un corpus d'interactions naturelles afin de vérifier si l'effet de fréquence relevé dans de précédentes études

au sujet de l'élision des schwas initiaux de polysyllabes est également perceptible en ce qui concerne la fréquence d'usage des collocations Clitique + X.

2 Facteurs influençant la (non) production du schwa : état des lieux

Avant d'entrer dans les détails de notre étude, il apparaît important de procéder à un état des lieux des travaux existants concernant les différents facteurs influençant le comportement du schwa. Nous n'entrerons pas dans les débats sur le statut phonologique et lexical du schwa[1] ou la formalisation de son alternance avec zéro. La principale raison est que les différents courants théoriques ayant abordé la question ont rarement pris en considération les particularités du contexte qui nous intéresse aujourd'hui. Nous ne présenterons donc pas les différentes positions défendues par les phonologues quant au statut phonologique et la représentation lexicale du schwa. Nous mettrons en revanche l'accent sur les facteurs qui semblent conditionner la (non) réalisation du schwa, en mettant en particulier l'accent sur le contexte monosyllabique. Dans cet objectif, cette section s'attardera dans un premier temps sur les facteurs dits segmentaux et suprasegmentaux avant de décrire les autres facteurs (sociolinguistiques, situationnels…).

2.1 Facteurs segmentaux et suprasegmentaux

L'un des facteurs segmentaux influençant la production du schwa les plus souvent cités correspond à ce que Grammont nomme la loi des trois consonnes (1894, cité par Laks, Durand, 2000). Concernant le schwa, la loi l'amènerait à se maintenir afin d'éviter la production d'une suite de trois consonnes. Ces remarques ont souvent été réinvesties au sein des approches phonologiques linéaires qui se sont particulièrement concentrées sur la nature des consonnes entourant le schwa (voir notamment Delattre, 1951 et Dell, 1973). Tout comme la nature des consonnes entourant le schwa, l'influence de la structure des syllabes voisines et de la position du schwa dans le groupe ou le mot prosodique (en initiale, en finale ou à l'intérieur) a largement été traitée (voir notamment Andreassen, 2013 ; Côté, Morrison, 2007 ; Delattre, 1951 ; Dell, 1973 ; Eychenne, 2006). Plus spécialement, concernant les clitiques, il s'avère que le schwa se maintient davantage après une consonne (*il passe par le toit*) qu'en début de groupe prosodique (*le toit est haut*) ou après une voyelle (*casse pas le toit*).

Refusant d'opposer les schwas finaux aux autres types de schwas en leur attribuant des statuts différents (épenthétiques et sous-jacents), Côté (2007) et Côté et Morrison (2007) proposent une analyse unifiée du schwa. Selon eux, son comportement serait prévisible uniquement en fonction de contraintes segmentales et prosodiques en surface. Au niveau suprasegmental, le schwa apparaitrait davantage au sein d'un mot prosodique qu'à ses frontières afin de renforcer la perceptibilité des consonnes internes des mots prosodiques, moins saillantes que celles qui les bornent. Pour les auteurs, il n'est donc pas nécessaire de postuler un schwa sous-jacent dans les représentations des clitiques puisque leur distribution paraît entièrement conditionnée par des facteurs (supra)segmentaux : « *it can be argued that [clitic] schwa should not be present in*

[1] Épenthétique, sous-jacent ou résultant d'un processus d'allomorphie ou d'allophonie ; pour une synthèse sur la question, voir notamment Côté et Morrison (2007).

phonological representations, following the principle of lexical economy that has served to exclude predictable information from lexical forms. » (Côté & Morrison, 2007, p. 182).

2.2 Facteurs sociolinguistiques et discursifs

Parmi les conditions non (supra)segmentales influençant le comportement du schwa, l'une des plus souvent citées regroupe les facteurs sociolinguistiques. Comme le relèvent Laks et Durand (2000), Grammont (1914) prenait déjà en considération leur influence. Conscient que la « langue des boulevards extérieurs » se distingue de l'oral de la bourgeoisie parisienne, l'auteur se restreint à décrire le schwa tel qu'il se comporte dans une « prononciation parisienne cultivée nettement distincte de la norme orthoépique de l'académie, de la Comédie Française ou du style oratoire professionnel qu'il juge incohérents » (Laks, Durand, 2000 : 35). L'opposition entre deux types de français parlé est encore plus nette chez Delattre (1951) qui, lorsqu'il analyse le comportement du schwa en syllabe interne, ne manque pas de relever que certaines variétés de français semblent déroger au patron général de la loi qu'il décrit. Concernant les schwas des clitiques, il semblerait qu'ils soient particulièrement sensibles aux facteurs sociolinguistiques (Hansen, 2000). Comparant deux corpus de productions recueillies en situation d'interview et de conversations informelles, l'auteure relève que seul le contexte clitique « présente un maintien plus fréquent du E caduc parmi les Parisiens adultes cultivés que parmi les Parisiens défavorisés de la même génération » (2000 : §28). Avec la classe sociale du locuteur, la variété régionale de français parlée influe également sur la distribution de la variable schwa observée au cours de discussions. Les travaux sur corpus de Eychenne (2006) et Andreassen (2013) ont notamment mis en lumière de nettes différences au niveau du comportement du schwa des clitiques. Par exemple, lorsque ces derniers sont produits en début de groupe intonatif, le schwa est alors élidé dans 9% des cas en français méridional contre 43% en français suisse et 71% en français de l'Ouest canadien. Le protocole mis en place par les chercheurs permet également de mettre en évidence l'effet de la situation de communication sur le comportement du schwa. Concernant les clitiques, Eychenne (2006) relève par exemple que les locuteurs du Languedoc produisent invariablement le schwa en contexte de lecture et ne relève aucun cas d'élision, même chez les locuteurs les plus créatifs élidant la voyelle dans des proportions supérieures à la moyenne du corpus en conversation libre.

2.3 Schwa et effet de fréquence

L'effet de la fréquence d'emploi du mot ou du groupe prosodique revient fréquemment dans les travaux visant à déterminer les facteurs gouvernant les usages du schwa (Dell, 1973 ; Eychenne, Pustka, 2006 ; Hansen, 1994 ; Racine, Grosjean, 2002 ; Racine, 2008). Reprenant les observations faites par Hansen (1994), Racine et Grosjean (2002) notent que ce facteur de fréquence correspond au principe décrit par Zipf (1949) selon lequel plus un mot est fréquent, plus il sera « court » et donc produit avec le moins de syllabes possibles. La chute du schwa, en réduisant d'une syllabe la longueur du mot, participe à ce principe. Dans leur travail sur l'élision du schwa en français de Suisse romande, Racine et Grosjean (2002) et Racine (2008) ont testé l'influence de sept facteurs[2] sur l'élision du schwa en syllabe initiale de noms polysyllabiques. Parmi ceux-ci,

[2] Ces facteurs sont : la fréquence lexicale, la fréquence de la prononciation, l'environnement consonantique, la force articulatoire, l'effacement précédent, la vitesse d'articulation et l'importance discursive.

les auteurs ont notamment cherché à tester le rôle de la fréquence lexicale. Afin d'estimer la fréquence d'usage des 66 substantifs de leur étude, les auteurs ont demandé à quatorze locuteurs d'assigner un score allant de 1 (très peu fréquent) à 7 (très fréquent) à chacun des items. Ce facteur a ensuite été comparé aux taux d'effacement relevés pour les 66 substantifs dans les productions de seize étudiantes suisses romandes qui avaient pour consigne de raconter une histoire. Le coefficient de corrélation calculé entre les indices de fréquence lexicale et les taux d'élision du schwa révèle une corrélation positive, « moyenne » mais significative ($r = 0.441$; $p < 0.01$). Autrement dit, il semble que les locuteurs soient sensibles à la fréquence de production d'un item : plus il a tendance à être perçu comme fréquent, plus les locuteurs procèderont à l'élision du schwa de la syllabe initiale. Ces études ont également permis de mettre en évidence que certaines variables, souvent présentées comme des facteurs importants de variation comme l'environnement consonantique ou la force articulatoire par exemple, ne semblent pas avoir d'effet significatif sur le comportement du schwa. En revanche, des facteurs liés à l'usage comme la force de la fréquence d'emploi, souvent citée mais jamais intégrée, à notre connaissance, à un formalisme phonologique, se révèlent fortement corrélés aux taux d'élision relevés par les auteurs. En effet, le coefficient de corrélation multiple calculé à partir de seulement trois facteurs dont deux liés à l'usage (fréquence lexicale, fréquence estimée de la prononciation et effacement précèdent) et du taux d'élision du schwa s'avère relativement élevé ($R = 0.775$; $p < 0.001$). En prenant en compte ces trois facteurs uniquement, les auteurs expliquent ainsi un peu moins de 57% de la variation relevée dans leur corpus.

Concernant la fréquence d'usage des clitiques, Eychenne et Pustka (2006) ont cherché à observer le comportement du schwa dans des séquences « *je* + Verbe » chez vingt locuteurs de l'Aveyron enregistrés en situation de discussion libre. Si, parmi les 856 collocations extraites de leur corpus de données, 297 contiennent le monosyllabe *je* avec un schwa élidé (soit environ 35% des cas), ce taux cache une variation importante en fonction du verbe utilisé. Alors que le taux d'élision du schwa du pronom combiné avec les six verbes les plus fréquents du corpus[3] oscille entre 42% et 56%, ce taux est de seulement 16% pour l'ensemble des autres verbes.

L'étude que nous présentons a pour objectif de vérifier si l'élision du schwa des clitiques est également conditionnée (au moins en partie) par la fréquence d'usage des items. Notre étude portant sur des clitiques, il ne semble pas pertinent de nous appuyer sur leur fréquence brute. En effet les clitiques, s'ils ont une fonction syntaxique propre, ne peuvent pas apparaître seuls dans un énoncé. Ces formes, le plus souvent atones, se combinent obligatoirement avec un autre élément auquel elles sont antéposées ou postposées. Nous avons donc souhaité observer si la fréquence des collocations Clitique + X avait un impact sur le taux de réalisation du schwa au sein de ces collocations. Notre travail, s'il se rapproche des études de Racine (2008) et de Racine et Grosjean (2002), diffère cependant au niveau de la méthodologie qui relève d'une approche entièrement écologique.

3 Méthodologie et hypothèse de recherche

Pour mener à bien nos analyses, nous avons choisi de nous appuyer sur le corpus ALIPE[4] (Liégeois et al., 2014), constitué à partir du recueil de trente heures d'interactions familiales. Au

[3] Ces verbes sont : *croire*, *être*, *penser*, *pouvoir*, *savoir* et *trouver*.

[4] Corpus du projet ALIPE (Acquisition de la Liaison et Interactions Parents-Enfant).

total, trois enfants et leurs parents respectifs ont été enregistrés pendant dix heures au cours de moments de la vie quotidienne tels que le bain, le jeu ou le repas. Afin d'employer la méthodologie la moins intrusive possible et de récolter les énoncés produits en situation naturelle, un simple enregistreur numérique a été confié aux parents. Ces derniers ont reçu pour seule consigne de procéder à environ une heure d'enregistrement quotidien pendant une semaine au cours de moments propices aux interactions parents-enfant. Afin d'étudier le développement langagier des jeunes sujets, la même procédure a été renouvelée sur au moins deux temps de recueils espacés d'environ huit mois. Si cette méthodologie a bien évidemment permis de récolter des énoncés enfantins ainsi que du Discours Adressé à l'Enfant (DAE), il s'avère qu'un peu plus de 20% de la totalité des graphies transcrites (soit 33 632 graphies sur environ 166 000, cf. Table 1) sont issues d'interactions entre adultes (ou Discours Adressé à l'Adulte, DAA).

	Couple 1	Couple 2	Couple 3	TOTAL
Graphies transcrites (DAA)	10 091	9 987	13 554	33 632
Collocations clitique + X extraites	874	1 014	1 266	3 154
Collocations sélectionnées	652	715	977	2 344

TABLE 1 : Couverture du corpus d'étude

Afin de rendre nos résultats comparables à ceux des études concernant les polysyllabes, nous avons souhaité restreindre nos analyses au seul DAA. En effet, il apparaît que les locuteurs adultes de notre corpus produisent des énoncés lexicalement moins riches en DAE qu'en DAA (Liégeois, 2014). De plus, nous avons pu observer dans de précédentes études que le DAE est modulé au niveau phonologique, dans le sens où nos sujets adultes élident significativement moins le schwa des clitiques en DAE qu'en DAA (Liégeois et al., 2012 ; Liégeois, 2014).

Dans le cadre de cette étude, nous traiterons des énoncés des adultes de notre corpus, en regroupant les productions par couple. Cette méthodologie se justifie selon nous par le fait que, pour chaque couple parental, les locuteurs possèdent un profil sociolinguistique similaire (cf. Table 2). De plus, nos précédents travaux ont pu montrer que les taux d'élision en DAA, pour chacun des couples de notre corpus, étaient significativement identiques à la fois entre les temps de recueil et entre homme et femme. Nos analyses s'appuieront donc sur un total de 2 344 contextes sélectionnés à partir des 3 154 collocations Clitiques + X extraites de nos données du DAA (cf. Table 1). L'exclusion de près de 800 contextes se justifie à deux niveaux. Tout d'abord, nous avons restreint nos analyses aux collocations recensées au moins cinq fois dans nos données, ce qui représente 70 collocations Clitiques + X différentes. De plus, nous avons également exclu de nos analyses les cas particuliers tels que les répétitions de clitiques (*je je je vais partir* par exemple) et les enchaînements de clitiques concernés par le schwa (*je le fais* par exemple).

Contrairement aux études précédemment citées, nous avons mis en place une méthodologie de recherche fondée uniquement sur un corpus d'interactions spontanées. Cette méthode révèle plusieurs limites. Tout d'abord, l'annotation de la (non) production du schwa s'est révélée complexe, les deux transcripteurs-annotateurs (experts) étant parfois en désaccord. Cependant, un calcul du Kappa de Cohen, indice d'accord inter-juges, indique un résultat de 0.835 (92% d'accord, niveau dit "très satisfaisant"). Si, après concertation, les annotateurs étaient toujours en désaccord, le contexte était alors doublement annoté (présence et absence de schwa) et exclu des données utiles aux analyses. Les difficultés liées à la transcription des discussions spontanées

nous ont également amenés à exclure de nos analyses l'ensemble des contextes pour lesquels la transcription était incertaine (production couverte par un bruit parasite par exemple). Si nous sommes conscients des limites d'une telle méthode d'annotation, des contraintes d'ordre temporelles nous ont empêché de procéder à une analyse acoustique des contextes annotés. Nous envisageons toutefois de procéder à l'analyse acoustique d'une sélection de données afin de vérifier si elle corrobore ou non l'annotation perceptive.

	Couple 1		Couple 2		Couple 3	
	Femme	Homme	Femme	Homme	Femme	Homme
Âge	32 ans	31 ans	38 ans	38 ans	31 ans	33 ans
Niveau d'étude (ISCE[5])	5	5	5	4	5	6
IPSE[6]	77	76	79	63	66	74
Département(s) de résidence (codes Insee[7])	63, 37, 75, 92	63, 37, 75	58	58	73, 39, 63	26, 39, 63

TABLE 2 : Profil sociolinguistique des locuteurs

Toutefois, notre méthodologie possède à nos yeux un avantage déterminant puisqu'elle nous permet d'éviter de nous appuyer sur des estimations de fréquence, comme le font par exemple Racine (2008) et Racine et Grosjean (2002). La couverture de notre corpus nous le permettant, nous avons souhaité nous appuyer sur la fréquence des collocations effectivement produites par nos sujets. Ainsi, nous chercherons à observer si plus une collocation Clitique + X est employée dans notre corpus d'analyse, plus le schwa du monosyllabe est élidé au sein de celle-ci. De ce fait, les deux mesures que nous utiliserons pour notre test statistique sont issues du même corpus de données, produites par les mêmes locuteurs. La section qui suit présente les résultats obtenus.

4 Résultats

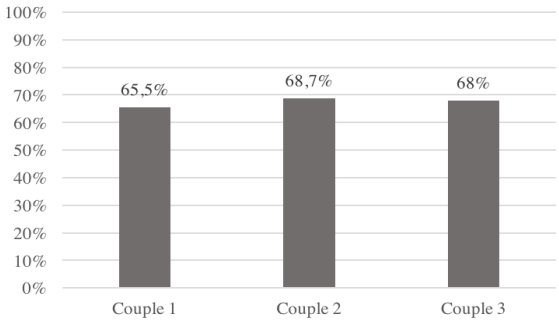
4.1 L'élision dans les productions des sujets adultes du corpus ALIPE

Les taux d'élision relevés dans les productions de chacun des couples de sujets sont très similaires, allant de 65,5% d'élision (couple 1) à 68,7% d'élision (couple 2, cf. Figure 1). Ainsi, notre corpus se révèle homogène, dans le sens où les usages de nos locuteurs ne diffèrent pas, ni à l'intérieur d'un même couple (Liégeois, 2014) ni entre les couples. Cette observation est confirmée par le calcul d'un Chi2 d'homogénéité qui ne révèle aucune différence significative entre les trois proportions relevées ($Chi2 = 1.7422$; $p > 0.05$).

5 ISCE : International Standard Classification of Education, défini par l'UNESCO.

6 IPSE : Indice de Position Socio-Économique (Genoud, 2011). Les locuteurs dont l'indice se situe entre 63 et 66 appartiennent à la catégorie dite « moyenne », tandis que les autres locuteurs dont les scores sont plus élevés appartiennent à la catégorie dite « moyenne-supérieure »).

7 Insee : Institut national de la statistique et des études économiques.



Nombre de collocations	652	715	977
Nombre de schwas élidés	427	491	664
Taux d'élision	65,5%	68,7%	68%

FIGURE 1 : Taux d'élision globaux

Selon nous, cette observation est due à deux facteurs principaux. Premièrement, nos sujets appartiennent à des catégories socioéconomiques proches et ne sont pas originaires de régions connues pour leur variété spécifique du français menant à des usages particuliers du schwa. De plus, les conditions d'enregistrement ont été similaires pour les trois couples : si l'objectif était de récolter principalement des interactions parents-enfant, nous avons pu recueillir des interactions entre adultes, essentiellement au cours des repas. Pendant ce temps de la vie quotidienne, il est en effet souvent arrivé que les discussions s'orientent autour de sujets auxquels les adultes seuls pouvaient participer (récits de la journée de travail ou discussions autour de questions d'actualité par exemple). Ainsi, nous disposons d'un ensemble de données cohérent que nous pouvons traiter dans son ensemble. Dans la suite de cette étude, les données parentales seront donc regroupées afin de tester l'hypothèse selon laquelle les taux d'élision du schwa des clitiques se révèleraient plus élevés au sein des collocations Clitique + X les plus fréquentes de notre corpus.

4.2 Effet de fréquence

Afin de vérifier notre hypothèse nous avons calculé, pour chacune des collocations produites au moins cinq fois par l'ensemble des locuteurs de notre corpus, deux mesures : la fréquence d'usage des collocations et le taux d'élision du schwa du clitique qu'elles contiennent. La mise en relation de la fréquence d'usage et des taux d'élision du schwa nous révèle une corrélation, certes faible, mais significative (Corrélation de Spearman : Rho = 0.3 ; $p < 0.012$). Bien que le coefficient de corrélation soit inferieur à celui relevé par Racine (2008) et Racine et Grosjean (2002), il montre selon nous un effet de la fréquence d'usage de la collocation sur le taux d'élision du schwa. De plus, nous relevons un effet significatif de la fréquence d'usage alors qu'aucun des autres facteurs influençant l'élision du schwa n'est contrôlé (nature des phonèmes entourant le schwa, contexte prosodique ou syntaxique par exemple). Selon nous, la faible valeur du coefficient de corrélation semble résulter de la grande variation du taux d'élision dans les collocations produites entre cinq et onze fois par les adultes. En effet, alors que certaines de ces collocations affichent un fort taux d'élision, comme *de place* par exemple (100% d'élision, 5/5), d'autres engendrent plus fréquemment un maintien du schwa, comme *de l'eau* par exemple (33,3% d'élision, 2/6). En

revanche, les collocations les plus fréquentes, c'est à dire produites plus de 20 fois par les locuteurs de notre corpus, font apparaître un usage plus homogène. En effet, pour chacune d'elle la variante non standard du monosyllabe est réalisée majoritairement, entre 54,2% des cas pour la collocation *me dit* et 96,2% pour la collocation *je peux*.

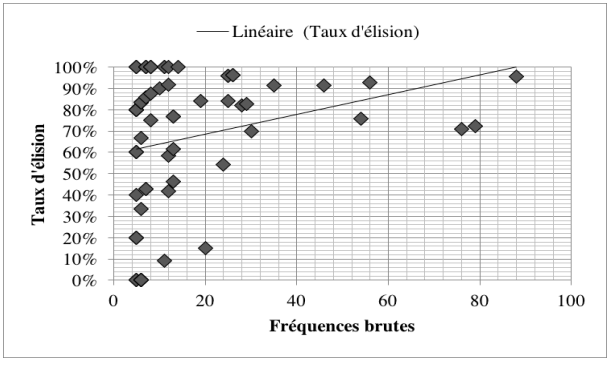


FIGURE 2 : Répartition des collocations Clitiques + X en fonction de leur fréquence d'usage (en abscisse) et du taux d'élision du schwa du clitique (en ordonnée).

5 Discussion

Nos données font donc apparaître un effet de la fréquence d'usage des collocations Clitique + X sur le taux d'élision du schwa. Sur ce point, les usages adultes diffèrent donc de ceux des enfants pré-lecteurs. En effet, comme nous avons pu le montrer dans une précédente étude, la fréquence d'usage des collocations Clitique + X par les jeunes locuteurs de notre corpus ne semble pas avoir d'incidence sur leurs usages du schwa (Liégeois, 2014). Concernant les données adultes, il semblerait en revanche que plus une collocation Clitique + X est fréquente en discussion et plus le schwa est souvent effacé en son sein. Ces résultats corroborent notamment ceux obtenus, essentiellement pour l'anglais, en ce qui concerne les liens entre la fréquence d'usage et l'effacement ou la réduction de consonnes et de voyelles internes ; pour une synthèse sur le sujet, voir notamment l'ouvrage coordonnée par Bybee et Hopper (2001).

Nous notons également que deux catégories de collocations semblent se dégager de nos données. D'un côté, les collocations présentes au maximum 20 fois dans notre corpus affichent une variabilité importante. D'un autre, les collocations produites plus de 20 fois par nos sujets affichent toutes un taux d'élision supérieur à 50%. Nos résultats issus d'un corpus d'interactions naturelles corroborent donc ceux obtenus par Racine (2008) et Racine et Grosjean (2002) concernant les schwas initiaux de substantifs polysyllabiques. Après avoir mis en évidence, dans de précédents travaux sur l'élision du schwa des clitiques, l'influence des propriétés phonologiques et phonétiques de l'environnement du schwa ainsi que celle du contexte de l'interaction (Liégeois, 2014), notre étude permet de relever un nouveau facteur lié à l'usage. Il conviendra, dans de prochains travaux, de tenter de relier l'ensemble de ces facteurs afin de proposer une étude, la plus complète possible, des conditions régissant le comportement du schwa des clitiques.

Références

ANDREASSEN, H. N. (2013). *Schwa: distribution and acquisition in light of Swiss French data*. University of Tromsø.

BYBEE, J., HOPPER, P. J. (dir.). (2001). *Frequency and the Emergence of Linguistics Structures* (TPS, Vol. 45). Amsterdam/Philadelphia : John Benjamins Publishing Company.

COTE, M. H., MORRISON, G. S. (2007). The nature of the schwa/zero alternation in French clitics: experimental and non-experimental evidence. *Journal of French Language Studies*, 17 (2), 159–186.

DELATTRE, P. (1951). Le jeu de l'e instable intérieur en français. *The French Review*, 24 (4), 341–351.

DELL, F. (1973). *Les règles et les sons*. Paris : Hermann.

EYCHENNE, J. (2006). *Aspects de la phonologie du schwa dans le français contemporain*. Université Toulouse-Le Mirail.

EYCHENNE, J., PUSTKA, E. (2006). The Initial Position in Southern French: Elision, Suppletion, Emergence. Actes des *JEL,* 199–204).

GENOUD, P. A. (2011). *Indice de position socioéconomique (IPSE) : un calcul simplifié*. Fribourg : Université de Fribourg.

GRAMMONT, M. (1894). *Le patois de la Franche-Montagne et en particulier de Damprichard (Franche-Comté). IV : La loi des trois consonnes*.

HANSEN, A. B. (1994). Etude du E caduc - stabilisation en cours et variations lexicales. *Journal of French Language Studies*, 4, 25–54.

HANSEN, A. B. (2000). Le E caduc interconsonantique en tant que variable sociolinguistique. *Linx*, 42, 45–58.

LAKS, B., DURAND, J. (2000). Relire les phonologues du français. Maurice Grammont et la loi des trois consonnes. *Langue Française*, 126 (1), 29–38.

LIEGEOIS, L. (2014). *Usage des variables phonologiques dans un corpus d'interactions naturelles parents-enfant : impact du bain linguistique et dispositifs cognitifs d'apprentissage*. Université Blaise Pascal.

LIEGEOIS, L., CHANIER, T., CHABANAL, D. (2014). *Corpus globaux ALIPE : Interactions parents-enfant annotées pour l'étude de la liaison*. Nancy : Ortolang.

LIEGEOIS, L., SADDOUR, I., CHABANAL, D. (2012). L'élision du schwa dans les interactions parents-enfant : étude de corpus. Actes des *29ème Journées d'Étude sur la Parole*, 313–320.

RACINE, I. (2008). *Les effets de l'effacement du Schwa sur la production et la perception de la parole en français*. Université de Genève.

RACINE, I., GROSJEAN, F. (2002). La production du E caduc facultatif est-elle prévisible ? Un début de réponse. *Journal of French Language Studies* 12 (3), 307–326.

ZIPF, G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge : Addison-Wesley.

Effort produit et ressenti selon le voisement en français

Camille Robieux, Thierry Legou, Yohann Meynadier, Christine Meunier
Aix Marseille Université, CNRS, Laboratoire Parole et Langage UMR 7309, 13100, Aix-en-Provence, France
camille.robieux@lpl-aix.fr

RÉSUMÉ

Les muscles laryngés et articulatoires sont impliqués dans la réalisation des traits qui distinguent les phonèmes. Cette étude porte sur l'auto-perception par les locuteurs et la répartition de l'effort vocal et articulatoire en fonction du trait de voisement en parole modale comparée à la parole chuchotée en français. Pour les 12 obstruantes du français, l'effort est ressenti plus important pour les voisées que les non voisées correspondantes, excepté dans le cas des fricatives labiodentales. Les analyses de la production des occlusives bilabiales montrent que l'effort laryngé est supérieur pour les consonnes voisées et l'effort articulatoire supérieur pour les non voisées, mais l'inverse pour les fricatives. Ces résultats indiquent que l'effort ressenti lors de sa propre production repose sur une perception prédominante de l'effort laryngé sur l'effort articulatoire en voix modale comme en voix chuchotée ; mais qu'il est cependant modulé selon le lieu et le mode d'articulation des consonnes.

ABSTRACT

Produced and perceived effort according to the voicing in French

Laryngeal and articulatory muscles are involved in the realization of the features that distinguish phonemes. This study focuses on self-perception by the speakers and repartition of vocal and articulatory effort according to the voicing in modal speech compared to whispered speech. For the 12 French obstruents, the perceived effort is greater for the voiced consonants than for the corresponding voiceless ones, except for the labiodental fricatives. The analysis of bilabial stops production show that laryngeal effort is greater for the voiced consonants and the articulatory effort greater for the voiceless ones, and reversely for the fricatives. These results indicate that perceived effort in his/her own production is based on a predominant perception of laryngeal effort over articulatory effort in modal speech as in whispered speech, but that it's modulated by the manner and the place of articulation.

MOTS-CLES : effort laryngé, effort articulatoire, obstruantes, voisement, parole modale, chuchotement

KEYWORDS: laryngeal effort, articulatory effort, obstruents, voicing, modal speech, whisper

1 Introduction

Cette étude porte sur l'effort produit et ressenti lors de la réalisation des consonnes obstruantes en français. L'effet du voisement et la répartition entre effort laryngé et effort articulatoire sont l'objet de cette recherche. La production des phonèmes requiert la tension ou la contraction de différents muscles : posturaux, respiratoires, mais aussi laryngés et articulatoires (Giovanni et al., 2014). Les muscles laryngés et articulatoires sont impliqués dans la réalisation des traits qui distinguent les

phonèmes entre eux. De ce fait, chaque phonème est produit avec une tension ou une contraction différente de mêmes muscles et donc avec un effort différent.

Les muscles laryngés sont impliqués dans la réalisation du trait de voisement. Dans cette étude, afin de comparer des phonèmes voisés et non voisés, nous avons choisi de nous concentrer sur les obstruantes du français /p, b, t, d, k, g, f, v, s, z, ʃ, ʒ/ pour lesquelles les muscles articulatoires sont impliqués dans la réalisation de deux autres traits : le mode et le lieu articulatoire.

Les obstruantes sont produites avec un effort musculaire laryngé plus ou moins important : les plis vocaux sont rapprochés pendant la production des consonnes, qu'elles soient voisées ou non voisées, mais l'effort laryngé, aussi appelé « effort vocal », est supposé plus important pour les consonnes voisées (Collier et al., 1979). En effet, pendant la réalisation de ces consonnes, la résistance laryngée au passage de l'air créée par l'adduction des plis vocaux doit être suffisamment importante pour permettre l'initiation et le maintien de la vibration des plis vocaux réalisant le voisement (Alipour et al., 1997 ;Alipour, Jaiswal, 2009 ; Rosenthal et al., 2014). Même en l'absence totale de vibration des plis vocaux, comme dans la parole chuchotée, l'adduction de ces plis vocaux reste effective (Sundberg, 2010 ; Crevier-Buchman, 2012). Même si la différence d'adduction entre les consonnes voisées et non voisées est réduite par rapport à la parole modale (Murry, 1976), elle semble rester distinctive en français (Meynadier, 2015).

Ces consonnes sont également produites avec un effort musculaire articulatoire plus ou moins important. Les articulateurs sont rapprochés pendant la production des obstruantes, qu'elles soient voisées ou non voisées, mais l'effort articulatoire est supposé plus important pour les consonnes non voisées (Marshal, 1983). En effet, pendant la réalisation de ces consonnes, la résistance articulatoire au passage de l'air doit être suffisamment importante pour créer un bruit de friction ou d'explosion. Cependant, la pression d'air appliquée derrière les articulateurs est plus forte pour les consonnes non voisées, pour lesquelles la résistance laryngée au passage de l'air pulmonaire est plus faible que pour les consonnes voisées (Netsell, 1969 ; Klitch, 1982). L'effort articulatoire peut également varier en fonction du mode et du lieu articulatoire (Malécot, 1966).

Cette étude porte principalement sur la répartition de l'effort de production des obstruantes en fonction du voisement. Dans l'hypothèse où l'effort laryngé primerait sur l'effort articulatoire, les obstruantes voisées seraient perçues comme nécessitant un effort de production plus important que les non voisées correspondantes. Au contraire, dans l'hypothèse où l'effort articulatoire primerait sur l'effort laryngé, les obstruantes non voisées seraient perçues comme nécessitant un effort de production plus important que les voisées correspondantes. Ces mêmes hypothèses seront également testées en parole non modale chuchotée, à savoir dans une parole nécessitant un moindre effort en production et une absence de vibration des plis vocaux, mais où le contraste de voisement semble pouvoir être maintenu en production, en français (Meynadier, 2015).

2 Expérimentation

2.1 Sujets

Les expérimentations ont porté sur trois groupes de sujets naïfs ne présentant ni trouble de la voix ni trouble de la parole, au total 114 sujets répartis en trois groupes. Chaque groupe comptait autant de sujets féminins que de sujets masculins. Le groupe *a* comptait 96 sujets âgés de 15 à 55 ans. Le groupe *b* comptait 24 sujets âgés de 41 à 55 ans, dont la moitié était issue du groupe *a*. Le groupe *c* comptait 12 sujets âgés de 23 à 32 ans, dont la moitié était également issue du groupe *a*.

2.2 Matériel

Cette étude porte principalement sur les consonnes labiales : occlusives bilabiales /p, b/ et fricatives labiodentales /f, v/, comparées en fonction du trait de voisement (voisée/ non voisée), du mode articulatoire (occlusive/ fricative) et du type de parole (modale/ non modale chuchotée). Sont également étudiées les occlusives apico-alvéolaires /t, d/ et dorso-vélaires /k, g/ ainsi que les fricatives apico-alvéolaires /s, z/ et post-alvéolaires /ʃ, ʒ/, mais uniquement en auto-perception ; les données aérodynamiques et physiologiques n'ayant pu être recueillies pour ces consonnes. Les consonnes, toujours associées à la voyelle /a/, ont été produites, selon les expériences, dans des syllabes CV, VC et VCV ou des trains de syllabes CV.

2.3 Méthode

Les deux premières expériences consistaient à produire des paires de syllabes CV, VC ou VCV en opposition de voisement (par ex., ap – ab, ava – afa). Elles ne faisaient pas l'objet d'enregistrements acoustiques. Après la production de chaque paire à voix haute, le sujet devait indiquer la syllabe la plus facile ou la plus difficile à produire, selon la consigne donnée, en fonction de ses propres sensations. Les paires étudiées ont été présentées dans les deux ordres inverses, par ex. ap – ab et ab – ap. Les 96 sujets du groupe *a* ont produit les 12 paires avec /p, b, f, v/ et les 24 paires avec /t, d, k, g, s, z, ʃ, ʒ/ en voix modale parmi 300 distracteurs. Les 24 sujets du groupe *b* ont produit les 12 paires labiales et les 24 paires alvéolaires, post-alvéolaires et vélaires en parole chuchotée parmi 180 distracteurs. Le laps de temps moyen écoulé entre les deux expériences était de 288 jours.

La troisième expérience ne portait que sur les consonnes /p, b, f, v/. Ici, des enregistrements ont été effectués. Les 12 sujets du groupe *c* devaient produire des trains de huit syllabes CV (par ex., pabafavapabafava). Seules les quatre syllabes centrales ont été analysées. Leur ordre a été randomisé par des carrés latins. Vingt-quatre trains de syllabes ont été produits par chaque sujet en parole modale puis, après sept autres types de production qui faisaient l'objet d'une expérience différente, en parole chuchotée. Pour les sujets ayant précédemment participé à une expérimentation d'auto-perception, le laps de temps moyen écoulé entre les deux expérimentations était de 368 jours.

Les données acoustiques, aérodynamiques et physiologiques ont été acquises avec EVA2. Les sujets étaient équipés d'un microphone fixé à une distance constante de la bouche. Ils étaient également équipés d'un tube inséré par la commissure des lèvres et maintenu entre les dents dont l'extrémité était placée au milieu de la cavité buccale, perpendiculairement au flux d'air, afin de mesurer la pression intra-orale ainsi que d'un masque placé à l'extrémité du conduit vocal afin de mesurer le débit d'air buccal. Les sujets portaient un pince-nez afin d'éviter toute déperdition d'air nasale. Ils étaient aussi équipés d'un capteur de pression de contact inséré dans une gaine fixée sur la gencive, en dessous les incisives inférieures, et sur le menton et placé au niveau de la lèvre inférieure afin d'enregistrer la pression des articulateurs, c'est-à-dire de la lèvre supérieure pour les occlusives bilabiales et des incisives supérieures pour les fricatives labiodentales.

Les mesures de la valeur de pression intra-orale maximale (PIO, en hPa) atteinte pendant la production de chaque consonne ainsi que le débit d'air buccal correspondant (DAB, en L/s) ont été faites sous le logiciel Phonedit (http://www.lpl-aix.fr/~lpldev/phonedit). Pour les occlusives, le pic de PIO survient le plus souvent au relâchement de la consonne et correspond donc à un DAB positif. A partir de ces valeurs, la résistance articulatoire au passage de l'air au niveau des lèvres (RA, en hPa/L/s) a été calculée en divisant la PIO par le DAB. Pour chaque consonne, les valeurs minimales et maximales de pression de contact ont été relevées afin de calculer l'empan maximal de pression mécanique aux lèvres lors de la consonne (en unité arbitraire). Enfin, dans chaque train de syllabes,

nous avons supposé la pression sous-glottique constante (Löqvist, 1975) et nous avons calculé la différence de résistance laryngée entre l'obstruante voisée et la non voisée correspondante (ΔRL, en hPa/L/s), en divisant la différence de PIO entre les deux consonnes par la différence de DAB. Les tests statistiques (ANOVA, test de Fisher) ont été réalisés sous Statview. Nous présentons la valeur de Fisher (F), le seuil de significativité (p) fixé à 5%, ainsi que la taille de l'effet (R^2).

3 Résultats

3.1 Auto-perception de l'effort de production

Les résultats des expériences d'auto-perception sont représentés sur la Figure 1. En parole modale, les résultats sont contraires pour les occlusives et les fricatives ($F(1,1150) = 60,9$; $p < 0,0001$; $R^2 = 5,0\%$). Pour les consonnes occlusives, les sujets ont perçu leur effort de production supérieur pour la voisée /b/ par rapport à la non voisée /p/ ($F(1,575) = 49$; $p < 0,0001$; $R^2 = 7,9\%$) alors que, pour les fricatives, ils ont ressenti un effort de production plus important pour la non voisée /f/ que pour la voisée /v/ ($F(1,575) = 16$; $p < 0,0001$; $R^2 = 2,8\%$).

En parole chuchotée, les résultats sont également contraires pour les occlusives et les fricatives ($F(1,286) = 5,1$; $p = 0,025$; $R^2 = 1,7\%$). Pour les occlusives, l'effort de production ressenti pour /b/ chuchoté est réduit par rapport à /b/ modal ($F(1,718) = 6,3$; $p = 0,013$; $R^2 = 0,8\%$). Il n'y a plus de différence significative entre la voisée /b/ et la non voisée /p/ ($F(1,143) = 0,4$; $p = 0,51$). Pour les consonnes fricatives, l'effort de production perçu n'est pas modifié par rapport à la parole modale ($F(1,718) = 0,2$; $p = 0,65$) mais, comme en parole modale, les sujets ont ressenti un effort supérieur pour la non voisée /f/ que pour la voisée /v/ ($F(1,143) = 6,5$; $p = 0,012$; $R^2 = 4,3\%$).

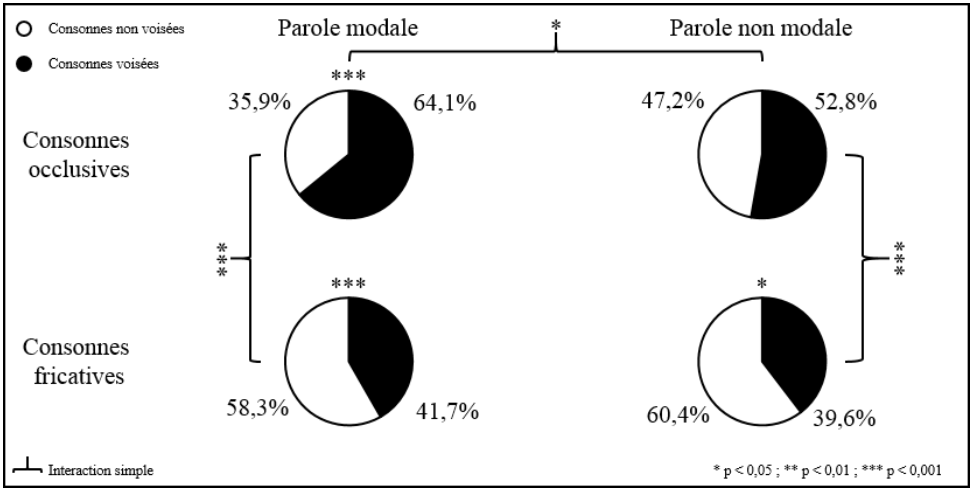


FIGURE 1 : Auto-perception de l'effort de production (%) pour les consonnes /p, b, f, v/

Les résultats de ces deux expériences pour les obstruantes alvéolaires, post-alvéolaires et vélaires sont présentés dans la Table 1. L'effort de production perçu pour les voisées /d, g, z, ʒ/ est supérieur à celui des non voisées correspondantes /t, k, s, ʃ/, en parole modale et en parole chuchotée. Ces résultats sont cohérents avec ceux obtenus pour les consonnes occlusives bilabiales en parole modale mais aussi en parole chuchotée, même si la différence entre voisée et non voisée n'était pas

statistiquement significative pour cette dernière. Seules les fricatives labiodentales présentent un schéma différent, peut-être du fait de la réduction de la sensibilité au niveau des dents.

Nous notons que l'effet du voisement est plus important en parole chuchotée qu'en parole modale pour les fricatives alvéolaires ($F(1,718) = 21,9$; $p < 0,0001$; $R^2 = 2,7$ %) et post-alvéolaires ($F(1,718) = 7,6$; $p = 0,006$; $R^2 = 0,9$ %) alors qu'il n'y a pas d'interaction significative pour les occlusives alvéolaires ($F(1,718) = 0,1$; $p = 0,7$) ou vélaires ($F(1,718) = 2,4$; $p = 0,13$).

Consonnes		Parole modale	Parole non modale
Occlusives	Apico-alvéolaires	t = 39,9% < d = 60,1% ***	t = 38,2% < d = 61,8% **
	Dorso-vélaires	k = 29,5% < g = 70,5% ***	k = 36,1% < g = 63,9% ***
Fricatives	Apico-alvéolaires	s = 41,7% < z = 58,3% ***	s =20,8% < z = 79,2% ***
	Post-alvéolaires	ʃ = 36,5% < ʒ = 63,5% ***	ʃ = 24,3% < ʒ = 75,7% ***

* p < 0,05 ; ** p < 0,01 ; *** p < 0,001

TABLE 1 : Auto-perception de l'effort de production (%) pour les consonnes /t, d, k, g, s, z, ʃ, ʒ/

3.2 Résistance articulatoire et pression de contact

Les résultats concernant les données aérodynamiques articulatoires sont représentées sur la Figure 2. Elles ne concernent que la production des obstuantes labiales. L'effort articulatoire de production, mesuré par la résistance articulatoire, est plus important pour les non voisées que pour les voisées en parole modale ($F(1,494) = 59,0$; $p < 0,0001$; $R^2 = 10,7\%$) et en parole non modale chuchotée ($F(1,525) = 29,1$; $p < 0,0001$; $R^2 = 5,2\%$) pour les occlusives, et seulement en parole chuchotée ($F(1,565) = 11,2$; $p = 0,0009$; $R^2 = 1,9\%$) pour les fricatives. Les fricatives en parole modale ne montrent pas de différence de résistance articulatoire selon leur voisement ($F(1,559) = 3,0$; $p = 0,083$) bien qu'elle semble plus importante pour les voisées que pour les non voisées.

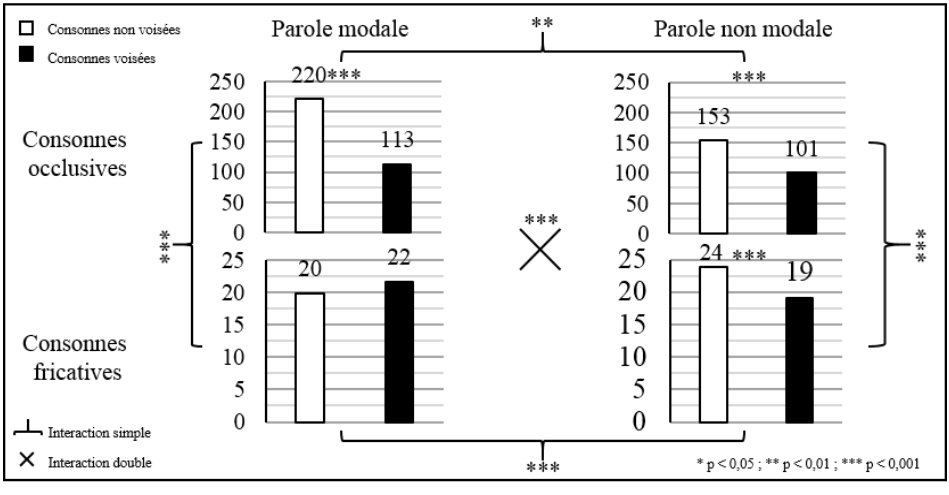


FIGURE 2 : Résistance articulatoire (hPa/L/s) pour les consonnes /p, b, f, v/

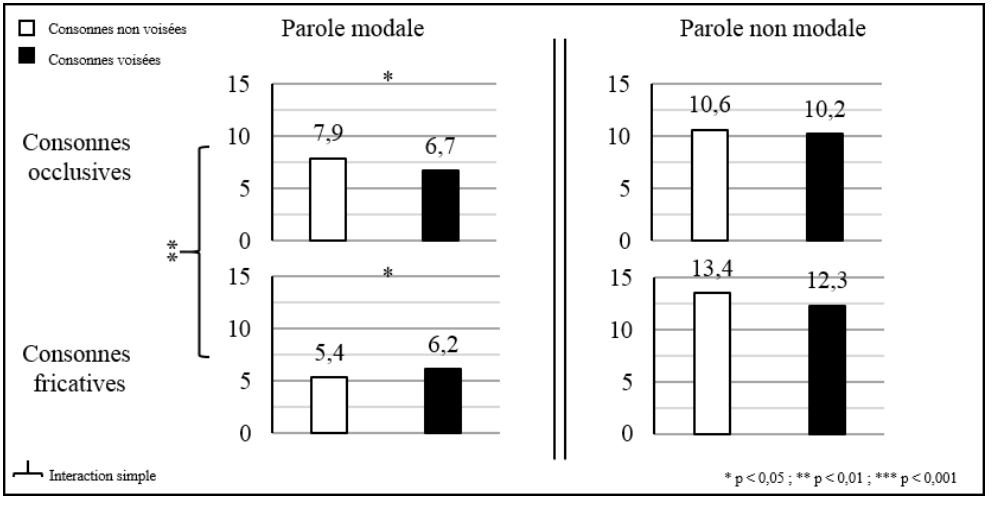


FIGURE 3 : Pression de contact (unité arbitraire) pour les consonnes /p, b, f, v/

Les données physiologiques issues de la production des obstruantes labiales, sont représentées sur la Figure 3. Concernant les mesures de pression de contact sur la lèvre inférieure, les résultats corroborent les comportements observés pour la résistance articulatoire aux lèvres. Obstruantes voisées et non voisée montrent une différence significative de pression de contact, uniquement en parole modale. Il existe une interaction croisée ($F(1,757) = 8,8$; $p = 0,003$; $R^2 = 1,1$ %) : l'occlusive non voisée /p/ est produite avec un contact plus fort que la voisée /b/ ($F(1,459) = 5,8$; $p = 0,016$; $R^2 = 1,3$), mais l'inverse est observé pour les fricatives : la voisée /v/ est articulée avec une pression des articulateurs plus forte que la non voisée /f/ ($F(1,298) = 5,3$; $p = 0,022$; $R^2 = 1,5$%).

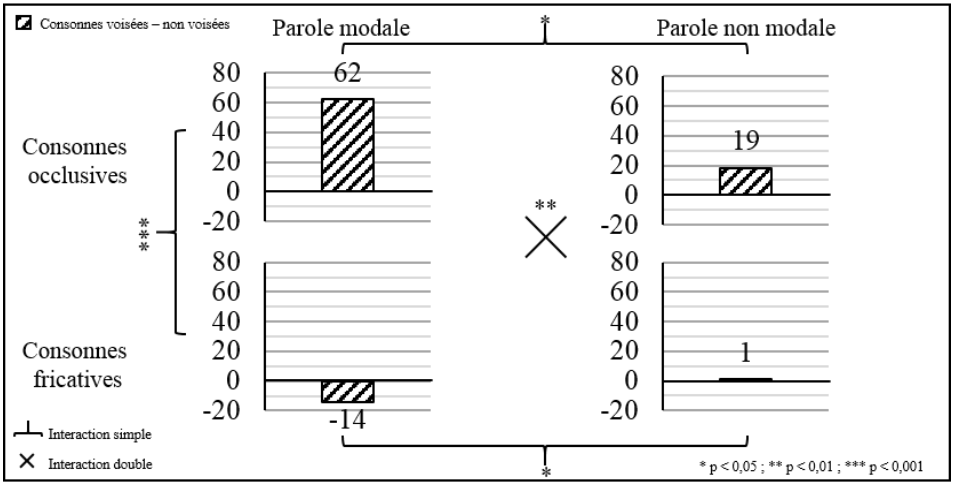


FIGURE 4 : Différence de résistance laryngée (hPa/L/s) pour les consonnes /p, b, f, v/

3.3 Différence de résistance laryngée

Les résultats concernant les données aérodynamiques laryngées sont représentées sur la Figure 4. Elles ne concernent là encore que la production des obstruantes labiales. Pour les occlusives, l'effort laryngé est supérieur pour les consonnes voisées /b/ que pour les non voisées /p/ en parole modale, avec une différence de 62 hPa/L/s, mais également en parole non modale, bien que la différence soit réduite à 19 hPa/L/s ($F(1,519) = 6,0$; $p = 0,015$; $R^2 = 1,1\%$).

Pour les fricatives, au contraire, l'effort laryngé est supérieur pour les consonnes non voisées /f/ que pour les consonnes voisées /v/ en parole modale, avec une différence de 14 hPa/L/s, mais cette différence est annulée en parole non modale chuchotée ($F(1,546) = 6,1$; $p = 0,014$; $R^2 = 1,1\%$).

4 Discussion

Dans cette étude, nous avons notamment mesuré la résistance articulatoire et la résistance laryngée au passage de l'air pendant la production d'obstruantes labiales. Ces mesures de résistance semblent être des indicateurs fiables de l'effort musculaire fourni pour atteindre et maintenir des positions articulatoires et laryngées permettant de réaliser les traits phonémiques. En effet, la mesure de résistance articulatoire est cohérente avec la mesure de force de contact réalisée. La combinaison des paramètres de pression et de débit avait, de plus, déjà été étudiée pour l'estimation de la constriction articulatoire (Pelorson, 2001). Quant à la mesure de résistance laryngée, elle est couramment utilisée pour évaluer l'effort vocal (Alipour et al., 2009 ; Rosenthal et al., 2014) et, dans notre étude, elle correspond à la réalisation du voisement. Nous discutons ici les liens entre l'effort produit et l'effort ressenti.

Pour les occlusives, en parole modale, bien que la consonne non voisée /p/ soit produite avec une résistance articulatoire deux fois supérieure à la consonne voisée /b/ ainsi qu'une pression de contact supérieure, cette dernière est tout de même perçue comme nécessitant un effort de production plus important car elle est produite avec une résistance laryngée supérieure. La résistance laryngée lors de la production de la consonne /b/ était en moyenne de 62 hPa/L/s supérieure à celle mesurée lors de la production de la consonne /p/. La résistance laryngée n'étant pas nulle lors de la production d'une consonne non voisée car les plis vocaux sont rapprochés (Crevier-Buchman, 2012 ; Meynadier, 2015), nous pouvons supposer que la résistance laryngée réelle lors de la production de la consonne /b/ est encore plus élevée que 62 hPa/L/s alors que la résistance laryngée moyenne mesurée indirectement lors de la production d'une voyelle /a/ en anglais se situe entre 40 et 50 hPa/L/s (Awan, 2013). L'effort laryngé serait donc plus important pendant la production d'une consonne voisée que pendant la production d'une voyelle. Cette donnée peut être prise en compte dans les études sur les troubles de parole avec troubles de la voix associés.

Les résultats obtenus pour les occlusives /p, b/ en parole modale semblent généralisables aux consonnes /t, d/, /s, z/, /k, g/ et /ʃ, ʒ/ pour lesquelles les sujets du groupe *a* ont également perçu un effort de production supérieur pour la voisée que pour la non voisé correspondante. La configuration articulatoire lors de la production d'une consonne voisée réduit le débit d'air oral et augmente donc la pression intra-orale. Alors, pour initier et soutenir la vibration des plis vocaux, la pression sous-glottique mais aussi la constriction laryngée, c'est-à-dire la résistance, augmentent (Collier, 1979).

En parole non modale chuchotée, la consonne occlusive non voisée /p/ est encore produite avec une résistance articulatoire supérieure à la consonne occlusive voisée /b/ mais cette différence est deux fois moins importante qu'en parole modale et il n'y a pas de différence de pression de contact des

lèvres. De même, la consonne /b/ est produite avec une résistance laryngée supérieure à la consonne /p/ mais la différence est trois fois moins importante qu'en parole modale. Ceci résulte en une compensation de l'effort articulatoire et de l'effort laryngé : les sujets n'ont pas perçu de différence d'effort dans la production des voisées et des non voisées en parole non modale chuchotée.

Ces résultats, quant à eux, ne semblent pas généralisables aux consonnes /t, d/, /s, z/, /k, g/ et /ʃ, ʒ/ chuchotée pour lesquelles les sujets du groupe *b* ont perçu un effort de production supérieur pour la voisée que pour la non voisé correspondante en parole non modale chuchotée. Nous pouvons supposer que, pour les consonnes occlusives /t, d/ et /k, g/, la différence d'effort laryngé reste suffisamment importante pour primer sur la différence d'effort articulatoire. En effet, ces consonnes sont produites avec des lieux articulatoires plus postérieurs que les consonnes bilabiales. La petite taille de la cavité buccale augmente la pression intra-orale et augmente donc la résistance laryngée nécessaire pour initier et soutenir la vibration des plis vocaux. Quant aux consonnes fricatives /s, z/ et /ʃ, ʒ/, pour lesquelles l'effet du voisement sur la perception de l'effort est plus important en parole chuchotée qu'en parole modale, il est difficile de formuler des hypothèses solides sans plus de données aérodynamiques.

Pour les fricatives, en parole modale, contrairement à ce qui est attendu, c'est la consonne voisée /v/ qui est produite avec une résistance articulatoire et une pression de contact des articulateurs supérieures à la non voisée /f/ et c'est la consonne non voisée /f/ qui est produite avec une résistance laryngée supérieure à la consonne voisée /v/. Ces résultats peuvent avoir été biaisés par la présence du capteur de pression de contact sur la lèvre inférieure gênant la réalisation des deux fricatives. Cependant, ces résultats sont cohérents avec les données d'effort ressenti : là encore, l'effort laryngé primerait sur l'effort articulatoire et ce serait donc la consonne non voisée /f/ qui serait perçue comme nécessitant un effort de production plus important que la consonne voisée /v/. Nous pouvons faire l'hypothèse que la configuration articulatoire de la consonne /v/, avec sa résistance faible au passage de l'air, est idéale pour la vibration des plis vocaux et que, au contraire, un effort laryngé est requis dans la production de la consonne non voisée /f/ pour empêcher cette vibration.

En parole non modale chuchotée, la consonne fricative non voisée /f/ est produite avec une résistance articulatoire supérieure à la consonne fricative voisée /v/ mais une résistance laryngée quasiment équivalente. Ceci explique la perception d'un effort de production plus important pour la consonne non voisée que pour la consonne voisée. En suivant l'hypothèse développée dans le paragraphe précédent, l'effort laryngé pour la production de la consonne /v/ en parole chuchotée augmenterait au même niveau que pour la consonne /f/ afin d'empêcher la vibration des plis vocaux.

5 Conclusion

L'effort musculaire laryngé prime sur l'effort musculaire articulatoire dans la perception de l'effort de production des consonnes. L'effort laryngé étant le plus souvent supérieur pour les consonnes voisées, celles-ci sont perçues comme nécessitant un plus grand effort de production que les non voisées correspondantes, même si ces dernières sont produites avec un effort articulatoire supérieur. Cependant, ces conclusions reposent principalement sur le cas des occlusives bilabiales et pourraient varier en fonction du mode articulatoire – nous avons d'ailleurs observé un comportement différent des fricatives labiodentales – et en fonction du lieu articulatoire. Les résultats présentés ici peuvent apporter un nouvel éclairage dans l'étude phonétique des troubles vocaux et articulatoires. Ils seront complétés prochainement par des mesures acoustiques de l'effort laryngé des 12 obstruantes, comme le taux de voisement de la consonne ou encore la fréquence fondamentale relative.

Références

ALIPOUR F., SCHERER R.C., FINNEGAN E. (1997). Pressure-flow relationships during phonation as a function of adduction. *Journal of Voice* 11, 187-194.

ALIPOUR F., JAISWAL S. (2009). Glottal airflow resistance in excised pig, sheep, and cow larynges. *Journal of Voice* 23, 40-50.

AWAN S.N., NOVALESKI C.K., YINGLING J.R. (2013). Test-Retest Reliability for Aerodynamic Measures of Voice. *Journal of Voice* 27(6), 674-684.

COLLIER R., LISKER L., HIROSE H., USHIJIMA T. (1979). Voicing in intervocalic stops and fricatives in Dutch. *Journal of Phonetics* 7, 357-373.

CREVIER-BUCHMAN L. (2012). *Phonétique clinique. Contribution à la compréhension de la voix et de la parole normale et pathologique*, p. 95. Thèse HDR. Aix-Marseille Université.

GIOVANNI A., LAGIER A., HENRICH N. (2014). Physiologie de la phonation. *EMC – Oto-rhino-laryngologie* 9(2), 1-15.

KLICH R.J. (1982). Effects of speech level and vowel context on intraoral air pressure in vocal and whispered speech. *Folia Phoniatrica et Logopaedica* 34(1), 33–40.

LÖQVIST A. (1975). A study of subglottal pressure during the production of Swedish stops. *Journal of Phonetics* 3, 175-189.

MALÉCOT A. (1966). Mechanical pressure as an Index of 'Force of Articulation'. *Phonetica* 14, 169-180.

MARSHAL A. (1983). The fortis-lenis distinction in stops. *Speech Communication* 2, 115-118.

MCHENRY M.A., KUNA S.T., MINTON J.T., VANOYE C.R. (1996). Comparison of direct and indirect calculations of laryngeal airway resistance in connected speech. *Journal of Voice* 10, 236-244.

MEYNADIER Y. (2015). Aerodynamic tool for phonology of voicing. *USB Proceedings of the 18th International Conference on Phonetic Sciences*, paper#0497. Glasgow.

NETSELL R. (1969). Subglottal and intraoral air pressures during the intervocalic contrast of /t/ and /d/. *Phonetica* 20(2-4), 68–73.

MURRY T., BROWN W.S. (1976). Peak intraoral air pressures in whispered stop consonants. *Journal of Phonetics* 4, 183-187.

PELORSON X. (2001). On the meaning and accuracy of the pressure-flow technique to determine constriction areas within the vocal tract. *Speech Communication* 35, 179-190.

ROSENTHAL A.L., LOWELL S.Y., COLTON R.H. (2014). Aerodynamic and acoustic features of vocal effort. *Journal of Voice* 28, 144-153.

SUNDBERG J. (2010). Whispering – A single subject study of glottal configurations and aerodynamics. *Journal of Voice* 24, 574-584.

Entraînements à la prosodie des questions ouvertes et fermées de l'anglais chez des apprenants francophones

Anne Guyot-Talbot, Karin Heidlmayr, Emmanuel Ferragne
CLILLAC-ARP EA 3967, Université Paris Diderot
5 rue Thomas Mann, 75013 Paris, France
anne.talbot@univ-paris-diderot.fr

RESUME

Des étudiants en anglais étaient invités à lire trois types de phrases : assertions, questions fermées et ouvertes. Ils étaient ensuite soumis à 3 sessions d'entraînements où ils devaient répéter des phrases interrogatives prononcées par une anglophone. Après chaque phrase, leur contour de F0 sur la syllabe portant le noyau intonatif ainsi que celui de la locutrice anglaise étaient affichés à l'écran. Ces sessions devaient leur permettre d'inférer une règle du système intonatif de l'anglais qui induit, par défaut, un contour montant pour les questions fermées et un contour descendant pour les questions ouvertes. Puis, une nouvelle séance d'enregistrements permettait de collecter des phrases à comparer au pré-test pour juger l'efficacité de l'entraînement. Les résultats montrent une réduction significative de la distance entre les contours mélodiques des apprenants du groupe test et ceux de la locutrice modèle entre pré-test et post-test, ce qui suggère un effet bénéfique de nos entraînements.

ABSTRACT

Prosodic training for French students of English on Wh- and yes-no questions.
French students of English were recruited for a reading task in a pre-test post-test design. Three types of sentences were recorded: assertion, yes-no questions and Wh- questions. The students then had to go through three training sessions over three days on the pronunciation of questions imitated from a native model. During the training, both native and non-native nuclear syllable pitch contours were displayed after each sentence utterance. The implicit training was expected to allow students to infer the default intonation patterns of both types of English questions: rising for yes-no questions, and falling for Wh- questions. A new set of sentences was recorded as a post-test to observe the effect of training. Results show a significant reduction of the distance between the pitch contours of the students in the test group and those of the native speaker between pre and post-test. Test group students seemed to have benefited from the training.

MOTS-CLES : Prosodie, F0, anglais langue étrangère.
KEYWORDS: Prosody, F0, English as a foreign language.

1 Introduction

Dans le domaine des Études Anglophones en France, le cursus typique inclut des cours destinés à améliorer la prononciation des étudiants. Il est donc convenu, au moins tacitement, que l'objectif est d'amener les étudiants à atteindre une précision articulatoire tendant vers celle de locuteurs natifs, précision articulatoire jugée superflue par la plupart des auteurs s'exprimant sur la question : ils

privilégient en effet d'autres compétences, telles que l'intelligibilité et la compréhensibilité (Munro et al., 2006 ; Thomson & Derwing, 2014). Il est pourtant avéré que parler avec un accent étranger entraîne une stigmatisation très préjudiciable aux locuteurs concernés (Gluszek & Dovidio, 2010).

Cet objectif de précision articulatoire (*nativeness*) s'appuie, à l'Université Paris Diderot, sur un enseignement explicite des règles qui régissent la correspondance entre graphèmes et phonèmes, des principes articulatoires de l'API, et des règles du placement de l'accent lexical. En revanche, l'intonation – la place de l'accent de phrase et le contour mélodique qui le matérialise – n'est pas abordée explicitement en première année. Dans cet article, nous nous concentrons sur l'intonation exclusivement.

Profitant de l'absence d'apport théorique sur la question, nous avons mené des séances d'entraînement intensif (répétition de phrases) à la production de deux contours intonatifs spécifiques (le ton montant, et le ton descendant) associés à deux structures syntaxiques (respectivement les questions fermées et les questions ouvertes) auprès d'étudiants de première année. L'entraînement avait 2 objectifs : 1) laisser inférer aux participants la règle du type de contour en fonction du type de question, et 2) favoriser la précision phonétique dans la production de ces contours par le biais d'un retour visuel représentant la courbe de F0 de l'apprenant et celle de la locutrice modèle cible. Concernant le premier point, Wells (2006) rappelle que, par défaut, le noyau d'une question fermée présente un contour montant, alors que celui d'une question ouverte comporte un contour descendant.

1.1 Intonation et langue seconde

Les étudiants impliqués dans l'étude sont des apprenants tardifs, qui ont appris l'anglais majoritairement à travers le système éducatif, c'est-à-dire, par le biais d'un enseignement explicite. Les études s'appuyant sur ce type d'enseignement impliquent généralement plusieurs heures d'instruction réparties sur plusieurs semaines (Ashby & Taniguchi, 2009 ; Atli & Bergil, 2012) ; et lorsque la notion enseignée n'est pas acquise, les auteurs attribuent cet échec à la durée trop courte de l'instruction (Dlaska & Krebeler 2013). Ici, nous tentons, au contraire, d'émuler les conditions naturelles d'acquisition d'une langue à travers des entraînements intensifs, brefs et peu nombreux qui devraient permettre de mettre au point plus rapidement la maîtrise phonétique de l'intonation qui manque aux apprenants (Jun et Oh, 2000).

Missaglia (1999) relève qu'acquérir ne serait-ce que des rudiments de traits prosodiques de la L2 suffit à atteindre un meilleur niveau de prononciation, tant au niveau accentuel ou intonatif, qu'au niveau des phonèmes. Elle suggère, au vu de ses résultats, que « l'accentuation et l'intonation ont une fonction de contrôle sur syllabes et segments », ce qui plaide en faveur d'une sensibilisation des apprenants plus tôt dans leur apprentissage.

Dans l'apprentissage d'une langue seconde, travailler à partir des caractéristiques de la langue maternelle est une pratique régulière (Pellegrino & Vigliano, 2015), qui permet une prise de conscience des différences entre L1 et L2. Ramírez Verdugo (2006) préconise même de mettre en avant les différences prosodiques, et notamment intonatives. Cette prise de conscience pourrait également concerner les similarités entre langues : pour nos étudiants, elle consisterait à rapprocher les schémas intonatifs attachés à chaque type de questions puisqu'ils présentent les mêmes contours attendus en français et en anglais (Hirst & Di Cristo, 1998)

Or au cours d'une précédente étude (Boissin et al., 2015), nous avions constaté chez des apprenants francophones de l'anglais, que le ton descendant, attendu par défaut sur les questions ouvertes, n'était pas maîtrisé (tons montants produits indifféremment sur les deux types de questions, ouvertes et fermées). C'est aussi le cas chez les apprenants chinois de l'anglais dans l'étude de Zhang et al. (2010). La production de ce ton s'améliore avec le niveau d'apprentissage (entre première et troisième année ; Boissin et al., 2015).

1.2 Mesures de similarité intonative

Comme l'explique Mennen (2015), alors qu'il est relativement aisé d'établir des prédictions s'appuyant sur des similitudes phonétiques et phonologiques entre les segments de la L1 et de la L2, comparer l'intonation des deux langues présente une difficulté accrue imputable à la complexité inhérente à l'intonation. Cette complexité découle en partie de l'interaction de l'intonation avec d'autres paramètres tels que l'accent de mot, le débit, la durée, etc. Elle résulte également de la nature moins « catégorielle » de l'intonation par rapport aux segments. Mennen (2015) encourage le recours au cadre phonologique autosegmental-métrique, qui consiste à catégoriser en un nombre fini d'étiquettes les contours intonatifs observés. Or, dans le cas d'apprenants d'une L2, imposer subjectivement une catégorie à des réalisations phonétiques qui pourraient varier de façon graduelle, à mi-chemin entre les normes de la L1 et de la L2 nous a paru quelque peu réducteur. Nous adoptons donc dans cet article une approche qui s'appuie sur une comparaison des données acoustiques.

Notre méthodologie s'inspire en partie de Rilliard et al. (2011), qui ont montré que la distance entre deux contours prosodiques, mesurée après déformation temporelle dynamique (*dynamic time warping* – DTW), constituait un bon indice de la distance linguistique entre ces deux contours. Le DTW permet de s'affranchir des différences de longueur des contours mais autorise également, contrairement à une simple normalisation du temps par interpolation, à des distorsions de phase plus locales. C'est donc sur la forme des contours que porte notre analyse, et non sur leur alignement avec les segments ou sur la hauteur absolue de F0, ces deux derniers types d'analyse étant beaucoup plus répandus dans la littérature que le premier (Post et al., 2007).

2 Expérience

2.1 Stimuli

L'expérience suivait un protocole prétest-entraînement-post-test. Le pré-test et le post-test contenaient chacun 54 phrases, différentes entres les deux tests, réparties en 3 types : assertives (*as* ; ex. *He was your mate*), questions fermées (*yn* ; ex. *Was he your mate?*) et questions ouvertes (*wh*; ex. *Who was your mate?*). L'accent nucléaire attendu dans chaque phrase était porté par un monosyllabe dont la fréquence d'occurrence dans le British National Corpus était d'au moins 800 (Kilgarriff, 1996). Le pré-test et le post-test étaient séparés par 3 séances d'entraînement, comportant chacune 60 phrases – uniquement des questions fermées ou ouvertes. Il s'agissait de 60 phrases différentes pour chaque entraînement. Ces phrases ont été construites sur le même principe que les phrases des tests.

2.2 Participants

Trente-cinq étudiants en première année à l'UFR d' Études Anglophones de l'Université Paris Diderot ont pris part à l'expérience sur la base du volontariat. Ils ont été répartis en 2 groupes : contrôle (16) et test (19). Une locutrice britannique, lectrice à l'UFR, servait de modèle.

2.3 Méthode

Lors d'une première séance, les participants des deux groupes enregistraient les 54 phrases du pré-test, qui étaient présentées l'une après l'autre sur un écran d'ordinateur avec le logiciel ROCme! (Ferragne et al., 2012). Le premier entraînement avait lieu juste après le pré-test. Les phrases du modèle étaient présentées auditivement et orthographiquement, et les participants étaient invités à répéter ce qu'ils venaient d'entendre. Ils appuyaient sur un bouton pour lancer l'enregistrement pour une durée fixe de 2 secondes à l'issue de laquelle le dernier mot était automatiquement extrait avec Praat, et le contour de F0 calculé sur ce mot, puis tracé, superposé au contour du locuteur-modèle. Après 3 séances d'entraînement sur 3 jours différents (répartis sur une semaine), les locuteurs étaient invités à enregistrer 54 nouvelles phrases en guise de post-test. Les entraînements étaient programmés à l'aide du *Demo Window* de Praat. Le groupe contrôle était invité à écouter des extraits d'émissions de radio en anglais pour une durée équivalente aux entraînements. Les enregistrements ont été effectuées dans une salle isolée acoustiquement avec un ordinateur équipé d'un microphone Audio-Technica AT 2020 ; le signal a été numérisé au formant PCM mono, 44,1 kHz, 16 bits.

3 Analyses et résultats

Avec le logiciel Praat, les syllabes nucléaires des 54 phrases × 2 tests (pré- et post-) × 35 locuteurs = 3780 phrases ont été segmentées manuellement. Les contours de F0 ont ensuite été estimés sur chacune de ces syllabes avec l'algorithme d'auto-corrélation de Praat. Le pas d'analyse était fixe (0,01 sec), et chaque contour a été inspecté visuellement et auditivement (en écoutant successivement le signal original et une resynthèse du F0 estimé) ce qui nous a donné l'opportunité de modifier les paramètres d'estimation dans le cas de mesures aberrantes. Les valeurs en Hertz ont été converties en demi-tons (par rapport à 1 Hz) et centrées sur 0 afin d'effacer les différences de hauteur moyenne propres à chaque individu. Les contours ainsi obtenus ont ensuite été comparés aux contours équivalents chez la locutrice modèle à travers plusieurs mesures de distances. Les analyses ont été effectués avec le logiciel R ; et en particulier, avec le package `dtw` (Giorgino, 2009). Nous avons effectué 3 types de mesures : 1) corrélation entre contours de l'apprenant et du modèle après normalisation temporelle par interpolation linéaire, 2) distance DTW entre contours de l'apprenant et du modèle, et 3) distance DTW entre les contours des questions ouvertes et ceux des questions fermées pour un même locuteur.

3.1 Corrélation après interpolation linéaire : apprenant – locuteur natif

Une ANOVA mixte incluant les facteurs Type de phrase (*as, wh, yn*), Séance (pré-test, post-test) et Groupe (test, contrôle) effectuée sur les coefficients de corrélation (transformés en arc sinus) entre la hauteur du son interpolée de l'apprenant et de la locutrice modèle, révèle une interaction significative entre les facteurs Type de phrase, Séance et Groupe ($F(2,3690)=4.48$, $p < .05$; FIGURE

1). Les analyses post-hoc montrent que dans le groupe test, il y a une augmentation significative de la corrélation pour les phrases *yn* ($p < .05$) ainsi que pour les phrases *wh* ($p < .001$) dans le post-test par rapport au pré-test, mais pas pour les phrases *as* ($p > .10$). Par contre, dans le groupe contrôle, les corrélations ne diffèrent entre pré- et post-test pour aucun type de phrases (ps > .10).

3.2 Distance DTW : apprenant – locuteur natif

Une ANOVA mixte incluant les facteurs Type de phrase, Séance et Groupe effectuée sur la distance de hauteur (données en demi-tons, centrées par locuteur) entre les contours des apprenants et ceux de la locutrice modèle obtenue par DTW montre une interaction significative entre les facteurs Type de phrase, Séance et Groupe ($F(2,3690)=8.02$, $p < .001$; FIGURE 2). Les analyses post-hoc montrent que dans le groupe test, il y a une diminution significative de la distance entre les contours de l'apprenant et ceux de la locutrice modèle pour les phrases *yn* ($p < .05$) dans le post-test par rapport au pré-test, mais pas pour les phrases *wh* et *as* (ps > .10). Par contre, dans le groupe contrôle, la distance augmente dans le post-test par rapport au pré-test pour les phrases *wh* ($p < .001$) mais ne diffère pas pour les phrases de type *yn* et *as* (ps > .10).

3.3 Distance DTW : apprenant *wh* - *yn*

Une ANOVA mixte incluant les facteurs Séance et Groupe effectuée sur la distance de hauteur (données en demi-tons, centrées par locuteur) entre les contours pour les phrases *wh* et *yn* des apprenants, obtenue par la méthode du DTW montre un effet principal du facteur Groupe ($F(1,35)=4.23$, $p<.05$), indiquant que la distance entre les phrases *wh* et *yn* est plus grande dans le groupe test (90.0 ± 63.8 demi-tons) par rapport au groupe contrôle (66.5 ± 55.5 demi-tons). En outre, il y a un effet principal du facteur Séance ($F(1,1208)=104.68$, $p<.001$), reflétant la taille de la distance entre les phrases *wh* et *yn* qui est plus grande dans le post-test (93.5 ± 71.9 demi-tons) que dans le prétest (65.0 ± 44.0 demi-tons). Toutefois, l'effet le plus important est l'interaction observée entre les facteurs Séance et Groupe ($F(1,1208)=18.84$, $p<.001$; FIGURE 3), qui montre que la distance entre les phrases *wh* et *yn* augmente de façon plus importante dans le groupe test ($t=10.78$, $p<.001$) que dans le groupe contrôle ($t=4.00$, $p<.001$).

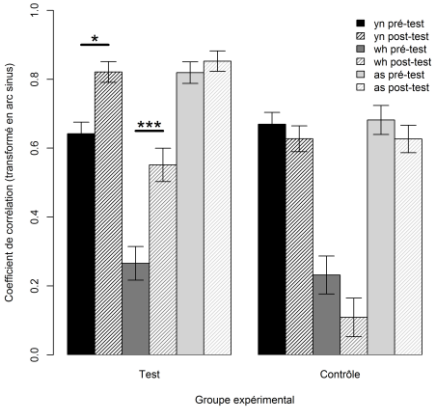


FIGURE 1 : Proximité (corrélation) après normalisation temporelle par interpolation entre contours d'apprenants et natifs

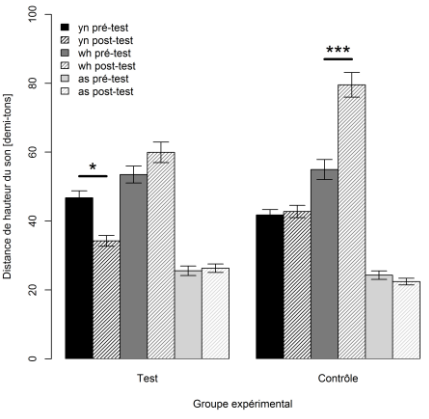


FIGURE 2 : Distance DTW entre contours d'apprenants et natifs

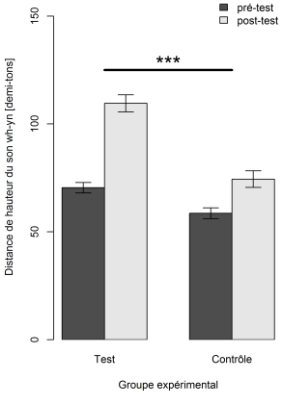


FIGURE 3 : Distance DTW entre questions *wh* et *yn* de chaque locuteur

4 Discussion et conclusion

Les résultats font donc apparaître un effet bénéfique de nos entraînements sur les questions ouvertes et sur les questions fermées, qui se traduit par une réduction de la distance – l'amélioration de la corrélation – entre les contours mélodiques des apprenants et ceux de la locutrice modèle – les contours étant normalisés temporellement par simple interpolation linéaire. L'effet n'est pas observé dans le groupe contrôle – qui ne suit pas d'entraînement, et n'émerge pas non plus pour les phrases assertives, qui ne font l'objet d'aucun entraînement. Il semble donc que les entraînements que nous

avons mis en place permettent aux apprenants d'améliorer leur production de l'intonation des questions ouvertes et fermées.

En s'affranchissant de la comparaison à la locutrice anglophone, et en se focalisant sur un éventuel accroissement de la distance DTW entre les contours des questions ouvertes et ceux des questions fermées entre pré-test et post-test, on observe une distance supérieure en post-test à la fois pour le groupe test et pour le groupe contrôle. L'interaction significative dans le modèle montre néanmoins que cette augmentation de la distance DTW entre pré- et post-test est significativement plus importante pour le groupe test. Notre effet d'entraînement reste donc robuste, et on peut éventuellement expliquer l'amélioration dans le groupe contrôle par un effet de pratique, traduisant une familiarisation des participants de ce groupe avec la tâche et le matériel.

Le schéma des analyses ayant la distance DTW entre apprenants et locutrice anglophone comme variable dépendante ne fait en revanche pas apparaître l'intégralité des effets escomptés, et est plus difficilement interprétable. En effet, si on note bien une amélioration – réduction de la distance DTW entre pré- et post-test – pour les questions fermées chez le groupe test (et pas chez le groupe contrôle), il n'est pas possible de mettre en évidence un effet bénéfique sur les questions ouvertes. En réalité, les performances du groupe contrôle se dégradent pour ce type de question, et il faut se contenter d'une absence de dégradation chez le groupe test. Les raisons d'un tel schéma restent encore obscures au stade où nous en sommes, mais un examen minutieux de chaque réalisation individuelle constitue une piste essentielle pour une future publication plus détaillée. Nous voyons en particulier deux facteurs qui pourraient conspirer à engendrer cette situation : 1) il est possible que, malgré le soin apporté à la conception des items de l'expérience et à l'estimation de la F0, les questions ouvertes du post-test présentent des caractéristiques phonétiques qui ont conduit à une moins bonne estimation des contours des apprenants (nous pensons par exemple à la difficulté constatée de mesurer F0 sur des voyelles brèves, à la nécessité de tronquer certains contours en raison d'erreurs d'estimation induites par les consonnes entourant la voyelle). 2) L'utilisation du DTW est peut-être particulièrement problématique pour comparer des contours d'apprenants et de natifs. En effet, pour obtenir des distances informatives dans notre cas, l'usage du DTW suppose que les séries temporelles à comparer aient une forme sous-jacente équivalente. Dans ce cas idéal, l'algorithme s'arrange avec les variations de débit de parole et les variations locales de phase. Dans notre cas, en revanche, les apprenants produisent peut-être des contours dont la forme globale est à mi-chemin entre contour montant et descendant, conduisant ainsi le DTW à générer des alignements aberrants. On peut également imaginer une plus grande variation inter-individuelle dans nos groupes pour les questions ouvertes, qui présentent précisément les contours pour lesquels, d'après notre expérience d'enseignant, les apprenants ont le plus de difficulté. Là encore, une inspection minutieuse des alignements obtenus pour chaque courbe constitue un prolongement naturel de cette étude.

Pour la suite, nous allons analyser ces mêmes contours avec la technique de *Functional Data Analysis*, qui a été récemment mise à profit pour étudier des enchaînements de deux voyelles en espagnol, permettant de distinguer les diphtongues des hiatus (Gubian et al. 2015). Cette technique comprend notamment une étape qui capture la variation de forme globale des contours à travers une extension de l'analyse en composantes principales appliquée aux contours lissés. En parallèle, afin d'évaluer la pertinence linguistique de toutes ces mesures, un panel de locuteurs natifs sera invité à écouter et à juger le degré d'amélioration de la production des apprenants entre pré- et post-test. Nous nous concentrons sur la forme des contours – plutôt que sur l'alignement de F0 avec les segments ou sur les différences de hauteur absolue – car cela nous paraît intuitivement très pertinent dans notre cas et, comme le notent Post et al. (2007), la forme des contours a souvent été négligée dans les études en prosodie par rapport aux deux autres paramètres que nous venons de mentionner.

Il sera néanmoins souhaitable d'examiner d'éventuelles différences d'alignement entre contours mélodiques et segments dans une analyse plus descriptive. S'appuyant sur la remarque de Post et al. (2007) qui note que la grammaire intonative de l'anglais est plus complexe que celle du français, on peut envisager l'enregistrement de quelques phrases équivalentes en français à titre de comparaison, et observer les conséquences acoustiques de cette différence.

Remerciements

Cette étude a bénéficié du soutien de l'IUF (E. Ferragne) et de l'Idex USPC (projet SOPHOCLE).

Références

ATLI I., BERGIL A. S. (2012). The effect of pronunciation instruction on students' overall speaking skills. *Procedia-Social and Behavioral Sciences* 46, 3665-3671.

ASHBY P., TANIGUCHI M. (2009). Assessing intonation. Actes de *Phonetics Teaching and Learning Conference*, 11-14.

BOISSIN J., GUYOT TALBOT A., FERRAGNE E. (2015). A cross-sectional acoustic study of L2 intonation patterns in 1st to 3rd year French students of English, *EuroSLA 25 "Second Language Acquisition : Implications for language sciences"*.

BOERSMA P., WEENINK D.(2016). Praat: Doing phonetics by computer (Version 6.0.13), Retrieved 01 31, 2016. Available from: <http://www.praat.org/>.

DLASKA A., KREKELER C. (2013). The short-term effects of individual corrective feedback on L2 pronunciation. *System*, 41, 25-37.

FERRAGNE E., FLAVIER S., FRESSARD C. (2012). ROCme! (Version 2.0) [Logiciel]. Consulté le 10 février 2016. *Téléchargeable à l'adresse :* www.ddl.ish-lyon.cnrs.fr/rocme

GIORGINO T. (2009). Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *Journal of Statistical Software*, 31(7), 1-24

GLUSZEK A., DOVIDIO J. F. (2010). The way they speak: a social psychological perspective on the stigma of non-native accents in communication. *Personality and Social Psychology Review*.

GUBIAN M., TORREIRA F., BOVES L. (2015). Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts. *Journal of Phonetics*, 49, 16-40.

HIRST D., DI CRISTO A. (1998). *Intonation systems: a survey of twenty languages*. Cambridge University Press.

JUN S. A., OH M. (2000). Acquisition of second language intonation. Actes de *INTERSPEECH*, 73-76.

KILGARRIFF A. (1996). BNC database and word frequency lists. Consulté le 10 février 2016. Téléchargeable à l'adresse : https://www.kilgarriff.co.uk/bnc-readme.html#lemmatised

LUTHY M. J. (1983). Nonnative speakers' perceptions of English "Nonlexical" Intonation Signals. Language Learning, 33(1), 19-36.

MENNEN I. (2015). Beyond Segments: Towards a L2 Intonation Learning Theory, in *Prosody and Language in Contact, L2 Acquisition, Attrition and Languages in Multilingual Situations.* Editors: Elisabeth Delais-Roussarie, Mathieu Avanzi, Sophie Herment, 171-188.

MISSAGLIA F. (1999). Contrastive prosody in SLA: An empirical study with Italian learners of German. Actes de ICPhS, 551-554. University of berkeley.

MUNRO M. J., DERWING T. M., SATO K. (2006). *Salient accents, covert attitudes: Consciousness-raising for pre-service second language teachers*. Prospect: an Australian journal of TESOL, 21(1), 65-77.

PELLEGRINO E., VIGLIANO D. (2015). *Self-imitation in prosody training: A study on Japanese learners of Italian*. Actes de *Workshop on Speech and Language Technology in Education, Satellite Event of INTERSPEECH 2015*, 53-57

POST B., D'IMPERIO M., GUSSENHOVEN C. (2007). Fine phonetic detail and intonational meaning. Actes de *International Congress of Phonetic Science* (ICPhS) 191-196.

RAMÍREZ VERDUGO D. (2006). A study of intonation awareness and learning in non-native speakers of English. *Language Awareness* 15(3), 141-159.

RILLIARD A., ALLAUZEN A., BOULA DE MAREÜIL P. (2011). Using Dynamic Time Warping to Compute Prosodic Similarity Measures. Actes de *INTERSPEECH*, 2021-2024.

THOMSON R. I., DERWING T. M. (2014). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, 1-20.

WELLS J. C. (2006). *English Intonation: An Introduction*. Cambridge University Press.

ZHANG S., LI K., LO W. K., MENG H. (2010). Perception of English suprasegmental features by non-native Chinese learners. Actes de *Speech Prosody 2010-Fifth International Conference.*

Estimation de la qualité d'un système de reconnaissance de la parole pour une tâche de compréhension

Olivier Galibert [1] Nathalie Camelin [2] Paul Deléglise [2] Sophie Rosset [3]
(1) LNE, F-78190 Trappes, France
(2) LIUM - Université du Maine, Le Mans, France
(3) LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
olivier.galibert@lne.fr, nathalie.camelin@lium.univ-lemans.fr,
paul.deleglise@lium.univ-lemans.fr, sophierosset@limsi.fr

RÉSUMÉ

Nous nous intéressons à l'évaluation de la qualité des systèmes de reconnaissance de la parole étant donné une tâche de compréhension. L'objectif de ce travail est de fournir un outil permettant la sélection d'un système de reconnaissance automatique de la parole le plus adapté pour un système de dialogue donné. Nous comparons ici différentes métriques, notamment le WER, NE-WER et ATENE métrique proposée récemment pour l'évaluation des systèmes de reconnaissance de la parole étant donné une tâche de reconnaissance d'entités nommées. Cette dernière métrique montrait une meilleure corrélation avec les résultats de la tâche globale que toutes les autres métriques testées. Nos mesures indiquent une très forte corrélation avec la mesure ATENE et une moins forte avec le WER.

ABSTRACT

Quality estimation of a Speech Recognition System for a Spoken Language Understanding task.

In this papier, we are interested in evaluating the quality of speech recognition system outputs considering a spoken language understanding task. The objective is to provide a tool that can select the most suitable automatic speech recognition system for a given dialogue system. We use different metrics in this study, including the WER, NE-WER and, ATENE. The latter metric was recently proposed for evaluating speech recognition systems considering a named entity recognition task. ATENE showed the better correlation with the results of the overall task among all other tested metrics. In this paper, the results indicate a stronger correlation between ATENE and the standard evaluation measure of spoken the language understanding task than with the other metrics.

MOTS-CLÉS : reconnaissance de la parole, compréhension, métrique d'évaluation.

KEYWORDS: Automatic Speech Recognition, Spoken Language Understanding, Evaluation Metric .

1 Introduction

Pouvoir estimer la qualité d'un système de Reconnaissance Automatique de la Parole (RAP) peut présenter un intérêt certain, notamment si l'on n'a pas accès au développement du système de RAP lui-même.

Actuellement, lorsqu'on mesure la qualité d'un système de RAP on utilise le plus souvent le taux

d'erreur mots (ou WER pour *Word Error Rate* (Pallett, 2003)). Or plusieurs travaux ont montré que le WER ne présentait pas toujours une bonne corrélation avec le score d'une tâche plus globale incluant la RAP. Par exemple si dans (Munteanu *et al.*, 2006), les auteurs observent, sur l'utilisabilité par des humains d'archives Web transcrites automatiquement, que la relation est linéaire entre le WER et la performance des humains à rechercher et trouver des informations dans les documents, ils observent aussi qu'en cas de grande différence de WER, celui-ci reste une métrique fiable. (Przybocki *et al.*, 1999) ont observé une forte corrélation entre le WER et la mesure utilisée pour l'évaluation de la détection des entités nommées (le SER, *Slot Error Rate*). Cette corrélation n'est pas du tout retrouvée dans une tâche similaire (voir (Ben Jannet *et al.*, 2015a)). Ceci dit, dans cette expérience, (Przybocki *et al.*, 1999) observent également que le système de RAP permettant d'aboutir au meilleur score global est celui qui obtient seulement la cinquième position en terme de WER.

Dans le contexte de la compréhension de la parole (ou SLU pour *Spoken Language Understanding*), des travaux (voir (Riccardi & Gorin, 1998; Wang *et al.*, 2003)) ont mis en évidence le fait que le WER n'est pas la métrique idéale pour évaluer la transcription pour la compréhension automatique de la parole. D'autres auteurs ont signalé des observations similaires sur les tâches de productions par des humains de résumés de séminaires préalablement transcrits automatiquement (Favre *et al.*, 2013). Dans le contexte des entités nommées, (Ben Jannet *et al.*, 2015a) ont proposé la métrique ATENE qui offre une meilleure corrélation que le WER par rapport au résultat global obtenu par des systèmes de Reconnaissance d'Entités Nommées (REN).

L'objectif de notre travail est, dans un premier temps, de voir si les résultats obtenus par ATENE sur une tâche de REN sont reproductibles sur une tâche de compréhension de parole ; dans un deuxième temps, nous souhaitons voir à quel point il est nécessaire ou pas de développer des modèles spécifiques étant donné que ces deux tâches (REN et SLU) sont conceptuellement proches.

L'article est organisé de la façon suivante : la section 2 présente quelques unes des métriques utilisées ou ayant été proposées pour l'évaluation de la qualité des systèmes de RAP. La section 3 présente le contexte de nos travaux et en particulier la tâche et les données utilisées. Puis la section 4.2 présente les expériences que nous avons menées pour répondre à nos deux questions ainsi que les différents résultats. Enfin la section 5 conclut ces travaux tout en ouvrant quelques perspectives.

2 Métriques d'évaluation utilisées en RAP

La métrique la plus utilisée pour évaluer la qualité d'un système de RAP est le WER. Cette métrique consiste à compter les erreurs selon les types prédéfinis que sont l'insertion, la suppression et la substitution déterminés par un alignement de Levenshtein entre la transcription manuelle (référence) et la transcription automatique (hypothèse). Plusieurs métriques alternatives existent. Certaines tentent de mesurer une perte d'information générale comme le RIL, *Relative Information Loss*, proposé par (Miller, 1955). Le RIL est fondé sur le principe d'information mutuelle et permet d'obtenir une mesure de la dépendance statistique entre le vocabulaire de la référence et celui de l'hypothèse. Cette mesure est représentée en termes d'entropie de Shannon. Par la suite, le WIL (Morris *et al.*, 2004) (*Word Information Loss*), qui est une approximation du RIL, a été introduit. Pour les taux d'erreurs élevés, le RIL et le WIL montrent des résultats intéressants (Morris *et al.*, 2004; McCowan *et al.*, 2004). Toujours dans le but de mesurer la perte d'information, (McCowan *et al.*, 2004) ont proposé d'adapter les métriques standards utilisées en extraction d'information, la précision (P), le rappel (R) et la f-mesure (F). L'idée général consiste à calculer le rappel et la précision au niveau des mots

en s'appuyant sur l'alignement entre la référence et l'hypothèse tel qu'il est produit par le calcul du WER. Deux autres métriques ont été proposées spécifiquement pour permettre d'évaluer la perte d'information étant donnée une tâche de reconnaissance d'entités nommées. La première métrique est le NE-WER (*Named Entity Word Error Rate*) proposé par (Garofolo *et al.*, 1999), qui mesure un WER dans les zones où la référence comprend une entité nommée. La seconde est ATENE (*Automatique Transcription Evaluation for Named Entities*) proposée par (Ben Jannet *et al.*, 2015b) qui est fondée sur un modèle probabiliste estimant le risque qu'une erreur de RAP induise une erreur de REN. Elle s'appuie sur une comparaison de probabilités de présence d'éléments d'intérêts (des entités dans le cas de la REN, des concepts si on l'utilise dans le cas de la compréhension par exemple) dans les transcriptions de référence et dans les hypothèses des systèmes de RAP. ATENE est la moyenne de deux mesures élémentaires $ATENE_{DS}$ et $ATENE_I$. Elle est donnée par l'equation 1.

$$ATENE = -100\frac{ATENE_{DS} + ATENE_I}{2} \tag{1}$$

$ATENE_{DS}$, donnée par l'équation 2, est une mesure du risque d'erreurs de suppression et de substitution d'entités engendré par les erreurs de RAP et $ATENE_I$, donnée par l'équation 3, est une mesure du risque d'erreurs d'insertion d'entités engendré par les erreurs de RAP.

$$ATENE_{DS} = \frac{\sum_{i=1}^{N} \Delta_p(\text{début}_i) + \Delta_p(\text{fin}_i)}{2N} \tag{2}$$

Δ_p est la différence de probabilités calculée sur des mots se trouvant en début et fin d'entités et N le nombre d'entités ou de concepts.

$$ATENE_I = \frac{\sum_{i=1}^{N_S} \Delta_{PS}(S_i)}{N_S} \tag{3}$$

Δ_{PS} est la différence de risque d'insertion calculée sur des segments de parole ne contenant pas d'entités nommées et N_S le nombre de segments entre entités ou concepts.

ATENE a obtenu de meilleures corrélations que le WER, le NE-WER ou encore les mesures de pertes d'information (WIL, et P, R, F) entre les performances obtenues par les systèmes de RAP et celles des systèmes de REN sur des données des campagnes QUAERO et ETAPE (Ben Jannet, 2015).

Comme nous l'avons dit, la tâche de REN semble relativement proche d'une tâche de compréhension de la parole, tout au moins telle que cette dernière est le plus souvent envisagée, c'est à dire comme une tâche de repérage de mentions de concepts dans un texte (voir la sous-section 3.1). Nous nous attendons donc à ce que ATENE permette d'obtenir une bonne corrélation avec l'évaluation globale de la tâche de compréhension. Ceci constitue la première partie des expériences que nous présentons ici. La deuxième partie, consiste à vérifier à quel point ATENE est dépendante d'un modèle correspondant strictement à la tâche. Pour cela nous tentons d'établir une correspondance entre la tâche de REN et celle de compréhension.

3 Tâche et Données

3.1 Compréhension de la parole

La tâche de compréhension de la parole consiste à associer un *sens* à un signal de parole. Donner un sens signifie que l'information, l'intention qu'a exprimée le locuteur doit pouvoir être traitée par l'ordinateur. Il s'agit donc de transformer le signal de parole en une interprétation sémantique, un langage formel qui traduit pour l'ordinateur le sens porté par les paroles du locuteur.

Généralement cette tâche s'exécute en deux temps. Tout d'abord, le signal de parole est transcrit automatiquement en une chaîne lexicale. Puis des systèmes de compréhension traitent cette chaîne lexicale afin d'en extraire une interprétation sémantique.

Le choix de la représentation sémantique est essentiel et doit être adapté aux données à analyser. On peut remarquer que dans la littérature, chaque tâche de compréhension adopte une représentation qui lui est propre. Cette diversité s'explique par la diversité des données traitées dans chaque application et par le fait qu'il n'existe pas de représentation sémantique générique qui puisse répondre aux besoins de toutes les applications. Par conséquent, dans la pratique, chaque tâche de compréhension adopte une représentation sémantique *ad hoc*.

Dans cet article, nous nous intéressons plus particulièrement aux applications de type dialogue homme-machine. Dans ce cadre, la tâche de compréhension consiste souvent à rechercher des renseignements qui sont liés à une base de données (tâche de *slot-filling*). Dans ce cadre particulier, chacun s'accorde notamment sur le fait que les éléments de base de la représentation sémantique sont des *concepts* et que chaque concept est associé à une *valeur* (e.g. : corpus ATIS (Hemphill *et al.*, 1990), corpus MEDIA (Bonneau-Maynard *et al.*, 2009), corpus PORTMEDIA (Lefèvre *et al.*, 2012)).

3.2 MEDIA : réservation d'hôtels et informations touristiques

Nous avons choisi de travailler sur le corpus MEDIA qui a été créé dans le but explicite de fournir à la communauté un corpus annoté sémantiquement afin d'évaluer et de comparer les différents systèmes de compréhension de la parole.

Le corpus comprend 1 257 dialogues entre un système simulé par un humain et un utilisateur souhaitant se renseigner ou réserver un hôtel. Le corpus a été transcrit et annoté manuellement. Les annotations sémantiques se composent de *modes*, *concepts*, *valeurs* et *spécifieurs*. Nous avons choisi dans un premier temps de ne tenir compte que du niveau *concept*, qui dans MEDIA correspond à 74 étiquettes définies selon une ontologie du domaine. Ils peuvent par exemple être des concepts généraux (réponse, nombre, temps-date, ...) ou se référer directement à des objets de la potentielle base de données (chambre, hotel). Chaque tour de parole est alors subdivisé en séquences de mots qui sont : soit associées à un concept (la séquence de mot est alors appelée support du concept) ; soit associées à l'étiquette *null* si la séquence de mots ne porte pas de sens vis à vis de l'application. Un exemple de texte annoté est donné dans la figure 1. Les 17 693 tours de paroles utilisateurs sont répartis selon trois sous-corpus : l'apprentissage contient 12 916 énoncés, le développement (DEV_MEDIA) en contient 1 259 et finalement le test (TEST_MEDIA) est composé de 3 518 énoncés. Un système de reconnaissance automatique de la parole identique à celui présenté dans (Bougares *et al.*, 2013) a été appliqué sur le corpus MEDIA afin d'obtenir des transcriptions automatiques. Il s'agit d'un système

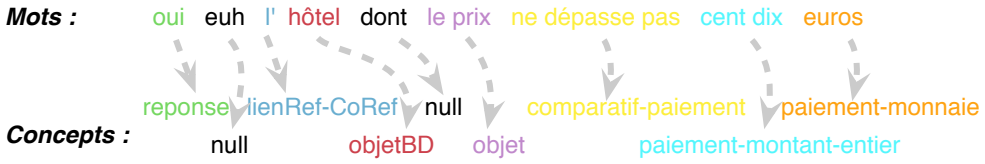


FIGURE 1 – Exemple d'interprétation sémantique en concept dans MEDIA

.

multi-passes fondé sur le projet CMU Sphinx [1], utilisant un modèle de langage quadrigramme adapté à MEDIA. Afin de générer différentes valeurs de WER, une segmentation manuelle ou automatique est utilisée et les paramètres du système ASR (poids du modèle de langage, poids d'insertion d'un mot, ...) ont été modulées. En conséquence, nous obtenons des transcriptions différentes qui présentent sur le DEV (4 systèmes) un WER oscillant entre 26,22% et 28,06% et sur le TEST (6 systèmes) entre 25,28% et 27,45%.

3.3 Implémentation du système de compréhension sur MEDIA

Dans (Hahn *et al.*, 2011), plusieurs systèmes de compréhension fondés sur différents algorithmes ont été implémentés et évalués. Les conclusions ont montré que le système fondé sur les champs conditionnels aléatoires (Conditional Random Field - CRF) (Lafferty *et al.*, 2001) était le plus performant. Récemment, plusieurs investigations de systèmes de compréhension de la parole fondés sur des réseaux de neurones ont été implémentés (*e.g.* (Mesnil *et al.*, 2015), (Liu & Lane, 2015), (Shi *et al.*, 2015)). Il a néanmoins été démontré dans (Vukotic *et al.*, 2015) que sur un corpus aussi complexe que MEDIA, le système fondé sur les CRF obtenait toujours les meilleurs résultats.

Nous avons donc choisi de ré-implémenter ce système en utilisant comme paramètres d'entrée : le mot, son étiquette sémantique, les 1 à 4 premières lettres du mot, les 1 à 4 dernières lettres du mot et également un paramètre indiquant l'absence ou la présence d'une majuscule sur la première lettre du mot. Pour obtenir différentes performances, nous utilisons plusieurs fichiers de paramètres afin que le CRF ne prenne pas en compte la totalité des paramètres d'entrée pour tous les systèmes mais seulement les 2 premiers, les 3 premiers... puis prenne en compte le contexte du mot plus ou moins élargi (0 à 2 mots avant, 0 à 2 mots après). Il en est de même pour le contexte de l'étiquette sémantique du mot. Cette dernière est issue d'un lexique obtenu manuellement. En effet dans MEDIA, le but du dialogue pour l'utilisateur est d'obtenir des informations qui sont stockées dans une base de données. Par conséquent, les noms de rues, de villes ou d'hôtels, les listes d'équipement de chambre, les types de nourriture, *etc.* sont connus. De plus, des mots plus généraux représentant les nombres, les jours, les mois sont également connus. Tous ces mots (spécifiques à la tâche SLU ou généraux) ont été rassemblés dans un lexique sémantique qui permet d'associer un mot à une étiquette sémantique.

1. http ://cmusphinx.sourceforge.net

4 Expériences

4.1 Méthodologie

Nous souhaitons dans un premier temps vérifier que nous pouvons utiliser la méthodologie proposée avec ATENE pour l'évaluation de système de RAP en vue d'une application de compréhension de la parole et obtenir de bonnes corrélations avec le CER [2], métrique utilisée ici pour évaluer la tâche de compréhension. Pour cela nous avons mis en place des expériences en suivant la méthodologie présentée dans (Ben Jannet, 2015). La première chose à faire est de développer les modèles nécessaires à l'utilisation de ATENE pour calculer les probabilités qui permettent d'estimer le risque d'erreurs. Un premier modèle permet d'estimer pour chaque mot sa probabilité d'être un début de support de concept-X ou non ; un deuxième permet d'estimer la probabilité pour un mot d'être une fin de support de concept-X ou non ; un dernier modèle permet d'estimer la probabilité qu'un mot soit n'importe où dans un support de concept ou non. Les deux premier modèles permettent de calculer $ATENE_{DS}$ en calculant le risque d'erreur induit par la RAP. Ce risque d'erreur est estimé par la différence entre les marges [3] calculées sur la transcription et la transcription automatique, comme indiqué dans l'équation 2. $ATENE_i$ est calculé en s'appuyant sur le troisième modèle qui estime la probabilité pour chaque mot d'être ou non dans un support de concept. Ce modèle n'est appliqué que sur les segments de parole hors concepts.

Nous avons appris différents modèles sur les données d'apprentissage du corpus MEDIA tel que présenté dans la section 3 : en utilisant l'ensemble des concepts détaillés ce qui représente 74 concepts y compris le concept NULL, en utilisant uniquement les têtes des concepts (par exemple COMMAND pour COMMAND-TACHE et COMMAND-DIAL) ce qui représente 22 concepts. Ceci représente six modèles qui s'appuient tous sur les mêmes traits : les mots dans une fenêtre $[-2, +2]$ et les préfixes et suffixes jusqu'à quatre caractères du mot courant.

Pour répondre à notre deuxième question sur la possibilité ou non d'utiliser des modèles appris sur une tâche différente, nous avons utilisé des modèles appris sur les données QUAERO [4] annotées en entités nommées. Nous n'avons considéré dans ces modèles que les entités pouvant potentiellement se rapprocher de certaines têtes de concept du modèle de compréhension : les dates (qui correspondent au concept TEMPS), les lieux désignant des villes (qui correspondent au concept LOCALISATION), les montants (qui correspondent aux concepts NOMBRE, SEJOUR, PAIEMENT et NOMBRENONDIGIT) , les organisations (qui correspondent aux concepts HOTEL et NOM) et les personnes.

Ici nous nous intéressons à la capacité des métriques d'évaluation des différents systèmes de RAP en fonction de la qualité de leur sortie étant donné la tâche de compréhension. Autrement dit, nous ne cherchons pas à dire tel système commet moins d'erreur que tel autre mais plutôt tel système comment moins d'erreurs ayant impact sur la tâche SLU que tel autre. Pour cela, nous comparons pour chaque métrique les classements des systèmes de RAP obtenus selon les performances de compréhension (donc selon le résultat obtenu en terme de CER) et les rangs des sytèmes de RAP selon les scores fournis par les métriques d'évaluation de RAP. Nous utilisons pour cela la corrélation de Kendall qui

2. Le Concept Error Rate, ou CER est une métrique d'évaluation largement utilisée en compréhension de la parole et qui s'apparente au WER. Dans le calcul du CER, on compare la liste des concepts de référence avec la liste des concepts fournis par le système de compréhension automatique. Il est à noter que le nombre de concepts dans une phrase n'est pas obligatoirement égal au nombre de mots.

3. La *Marge* est l'écart de probabilités entre le label attendu et le meilleur label pour un modèle donné (Ben Jannet, 2015).

4. Nous utilisons ces données car elles sont disponibles gratuitement pour la recherche académique, voir `http://catalog.elra.info/product_info.php?products_id=1195`.

reflète le degré de concordance et de discordance entre les rangs de deux classements. Cette mesure donne des valeurs comprises entre -1 et +1 dont la valeur absolue indique la puissance de corrélation entre les deux variables testées. Il est à noter que les corrélations ne sont pas calculables quand une des mesures comparées ne fournit que des scores identiques. Hors cela arrive parfois, en particulier avec CER et WER-NE. Nous nous sommes donc restreints aux dialogues où cette situation ne se produit pas.

Métrique	Dev.				Test.			
	Kendall	IC. 95%	#ASR	#Tests	Kendall	IC. 95%	#ASR	#Tests
WER	0,853	[0,598 ;1]	4	294	0,837	[0,603 ;1]	6	939
1-F	0,869	[0,641 ;1]	4	294	0,851	[0,656 ;1]	6	939
WER-NE	0,771	[0,425 ;1]	4	294	0,841	[0,545 ;1]	6	939
ATENE-SLU	0,885	[0,701 ;1]	4	294	0,858	[0,661 ;1]	6	939
ATENE-TOP	0,866	[0,612 ;1]	4	294	0,857	[0,661 ;1]	6	939
ATENE-NE	0,852	[0,563 ;1]	4	294	0,840	[0,613 ;1]	6	939

FIGURE 2 – Corrélation de Kendall moyenne avec intervalle de confiance à 95% entre les différentes métriques et le CER. #ASR indique le nombre de systèmes ASR comparés, #Tests indique le nombre de dialogues sur lesquels la corrélation de rang est calculée. ATENE-SLU réfère à la métrique ATENE utilisant les modèles appris sur les données d'apprentissage MEDIA de la tâche SLU, ATENE-TOP est identique mais utilise uniquement les têtes de concept et ATENE-NE renvoie à l'utilisation des modèles appris sur le corpus QUAERO pour la tâche NER.

4.2 Résultats

Le tableau 2 montre les différents résultats obtenus. Il contient les coefficients de corrélation (τ de Kendall) associés aux intervalles de confiance à 95 %. Les résultats sont donnés pour le corpus de développement et le corpus de test. Nous indiquons ce qui a été obtenu pour chacune des métriques évaluées : le WER, ATENE-SLU, c'est à dire ATENE construit avec les données adaptées à la tâche SLU, les données MEDIA, ATENE-TOP qui s'appuie sur les têtes de concept, ATENE-NE, construit avec les données du corpus QUAERO mais uniquement certaines entités, la F-mesure ou plutôt son complémentaire (1-F) et le NE-WER.

On constate que ATENE-SLU a effectivement une légèrement meilleure corrélation que WER. La différence est loin d'être aussi importante que ce qu'elle était sur les tâches d'EN telles que rapportées dans (Ben Jannet, 2015) mais reste présente. Il est intéressant de noter que l'intervalle de confiance est mieux resserré avec ATENE-SLU qu'avec le WER, indiquant que ATENE-SLU est moins bruitée, et donc probablement plus utile dans le cadre d'un développement de système de RAP avec une faible quantité de données de développement disponibles. On peut voir que la F-mesure est de façon inattendu un meilleur estimateur de la qualité des systèmes de RAP que le WER mais n'atteint pas le niveau d'ATENE. L'approche WER-NE ne semble pas produire de résultats particulièrement intéressants. La non-prise en compte des effets induits par les insertions de mot par les systèmes de RAP la rend tout aussi peu fiable que dans le cadre de la REN. La légère meilleure performance de ATENE-SLU par rapport à ATENE-TOP semble indiquer qu'ATENE a besoin de modèles relativement précis plutôt que larges. ATENE-NE confirme cette tendance. En effet, des modèles non spécifiques à la tâche ne permettent pas d'obtenir une corrélation raisonnable.

5 Conclusion et perspectives

Nous avons présenté une série d'expériences menées pour vérifier à quel point une métrique d'évaluation de systèmes de RAP peut corréler avec le résultat global d'une tâche de SLU. Nous avons comparé différentes métriques dont la métrique la plus utilisée, le taux d'erreurs mots (WER), et ATENE, une métrique récemment proposée pour l'évaluation de systèmes de RAP en vue d'une tâche de reconnaissance d'entités nommées. Cette dernière métrique a été implémentée d'une part en utilisant les données liées à la tâche considérée (les données MEDIA) mais aussi en utilisant les données liées à une tâche de reconnaissance d'entités nommées (les données QUAERO). Ce dernier cas avait pour objectif de mesurer la dépendance de ATENE aux données de la tâche. ATENE, dans sa forme la plus complète et la plus adaptée à la tâche, donne le meilleur résultat, c'est à dire un coefficient de corrélation supérieur à 0,85 avec un intervalle de confiance plus resserré qu'avec le WER. Nous allons poursuivre nos travaux dans deux directions. La première utilisation d'ATENE peut être de l'utiliser pour le *tuning* d'un système de RAP. Toutefois, nous considérons qu'améliorer ATENE, ou n'importe quelle autre métrique, pour étudier la corrélation entre les performances de RAP et celles de SLU passe par la définition d'une meilleure métrique SLU qui prenne en compte la totalité du problème, non seulement la détection des concepts mais aussi de leur mode et leur valeur.

Remerciements

Ce travail a été financé partiellement par le projet VERA - ANR 12 BS02 006 04.

Références

BEN JANNET M. A. (2015). *Évaluation adaptative des systèmes de transcription en contextes applicatifs*. PhD thesis, Université Paris Sud.

BEN JANNET M. A., GALIBERT O., ADDA-DECKER M. & ROSSET S. (2015a). How to evaluate asr output for named entity recognition ? In *16th Annual Conference of the International Speech Communication Association (Interspeech'15)*.

BEN JANNET M. A., GALIBERT O., ADDA-DECKER M. & ROSSET S. (2015b). How to evaluate asr output for named entity recognition ? In *Interspeech*, Dresden, Germany.

BONNEAU-MAYNARD H., QUIGNARD M. & DENIS A. (2009). Media : a semantically annotated corpus of task oriented dialogs in french. *Language Resources and Evaluation*, **43**(4), 329–354.

BOUGARES F., DELÉGLISE P., ESTEVE Y. & ROUVIER M. (2013). Lium asr system for etape french evaluation campaign : experiments on system combination using open-source recognizers. In *Text, Speech, and Dialogue*, p. 319–326 : Springer.

FAVRE B., CHEUNG K., KAZEMIAN S., LEE A., LIU Y., MUNTEANU C., NENKOVA A., OCHEI D., PENN G., TRATZ S., VOSS C. & ZELLER F. (2013). Automatic Human Utility Evaluation of ASR Systems : Does WER Really Predict Performance ? In *Interspeech, Lyon (France)*.

GAROFOLO J. S., VOORHEES E. M., AUZANNE C. G., STANFORD V. M. & LUND B. A. (1999). 1998 trec-7 spoken document retrieval track overview and results. In *Broadcast News Workshop'99 Proceedings*, p. 215 : Morgan Kaufmann Pub.

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. *Audio, Speech, and Language Processing, IEEE Transactions on*, **19**(6), 1569–1583.

HEMPHILL C. T., GODFREY J. J. & DODDINGTON G. R. (1990). The atis spoken language systems pilot corpus. In *Proceedings of the DARPA speech and natural language workshop*, p. 96–101.

LAFFERTY J. D., MCCALLUM J. D. & PEREIRA F. C. N. (2001). Conditional random fields : probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML-2001)*, San Francisco, CA, USA.

LEFÈVRE F., MOSTEFA D., BESACIER L., ESTÈVE Y., QUIGNARD M., CAMELIN N., FAVRE B., JABAIAN B. & ROJAS-BARAHONA L. (2012). Robustesse et portabilités multilingue et multi-domaines des systèmes de compréhension de la parole : les corpus du projet portmedia. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1 : JEP*, p. 779–786 : ATALA/AFCP.

LIU B. & LANE I. (2015). Recurrent neural network structured output prediction for spoken language understanding. In *Proc. NIPS Workshop on Machine Learning for Spoken Language Understanding and Interactions*.

MCCOWAN I. A., MOORE D., DINES J., GATICA-PEREZ D., FLYNN M., WELLNER P. & BOURLARD H. (2004). *On the use of information retrieval measures for speech recognition evaluation*. Rapport interne, IDIAP.

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. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, **23**(3), 530–539.

MILLER G. A. (1955). Note on the bias of information estimates. *Information theory in psychology : Problems and methods*, **2**, 95–100.

MORRIS A. C., MAIER V. & GREEN P. (2004). From wer and ril to mer and wil : improved evaluation measures for connected speech recognition. In *INTERSPEECH*.

MUNTEANU C., BAECKER R., PENN G., TOMS E. & JAMES D. (2006). The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, p. 493–502 : ACM.

PALLETT D. S. (2003). A look at nist's benchmark asr tests : past, present, and future. In *ASRU'03*.

PRZYBOCKI M. A., FISCUS J. G., GAROFOLO J. S. & PALLETT D. S. (1999). 1998 hub-4 information extraction evaluation. In *Proc. DARPA Broadcast News Workshop,(Herndon, Va, USA)*, p. 13–18.

RICCARDI G. & GORIN A. L. (1998). Stochastic language models for speech recognition and understanding. In *ICSLP*.

SHI Y., YAO K., CHEN H., PAN Y.-C., HWANG M.-Y. & PENG B. (2015). Contextual spoken language understanding using recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, p. 5271–5275 : IEEE.

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*.

WANG Y.-Y., ACERO A. & CHELBA C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. In *ASRU'03*, p. 577–582 : IEEE.

Etude acoustique du discours politique d'hispanophones : le cas de Hugo Chávez et de José Zapatero

Carmen Patricia Pérez
Université Paris Diderot, CLILLAC-ARP, 5 rue Thomas Mann, 75013 Paris, France
perez.patriccia@gmail.com

RESUME

Les styles de discours des hommes politiques peuvent être identifiés grâce à leurs réalisations prosodiques. On peut reconnaître un homme politique 'révolutionnaire' ou 'traditionnel' en écoutant quelques minutes de discours. Je me propose de montrer quels sont les paramètres prosodiques pertinents dans cette distinction en comparant les phonostyles de Hugo Chávez et José Zapatero. Je présente également le changement de phonostyle de Chávez dans deux situations différentes (c.-à-d. deux phono-genres), en interview et en public. Le modèle de Ph. Martin *Contraste de Pente Mélodique* est utilisé pour décrire la structure prosodique. Les analyses acoustiques montrent que les phonostyles de ces personnalités se différencient, dans le même phono-genre, dans la réalisation des contours de continuation, l'étendue du registre et le débit, alors que la construction des groupes intonatifs est semblable. Une brève étude sur les imitateurs de Chávez et de Zapatero est rajoutée pour montrer qu'ils reproduisent avec efficacité les paramètres acoustiques pertinents de ces leaders.

ABSTRACT

Politicians' speech styles can be distinguished thanks to their prosodic realizations. Generally, we can recognize a 'revolutionary' or a 'traditional' politician just listening to a few minutes speech; I propose to show which prosodic features enable us to do so, comparing Hugo Chávez and José Zapatero's respective phonostyles in public speeches. Moreover, I will show the differences between Chávez's own phonostyles according to the situation, interview and public speech ('phono-genres'). Ph. Martin's *Melodic Slope Contrast* model is used to describe the prosodic structure. The acoustic analysis shows that the phonostyle of these political leaders differs in the same 'phono-genre', mainly in the realization of continuation contours, the range and the speech rate, while the construction of the intonation phrases is the same. A short study of imitators' production has been added to show how they select the right and pertinent prosodic features of these leaders.

MOTS-CLES : intonation, discours public, discours politique, phonostyles, Chávez, Zapatero.
KEYWORDS: intonation, public speech, political speech, phonostyles, Chávez, Zapatero.

1 Introduction

Les hommes politiques sont souvent reconnaissables grâce aux réalisations prosodiques de leurs discours qui les rendent plus au moins charismatiques, intéressants et populaires. Ces réalisations se caractérisent par des événements prosodiques comme les proéminences de la fréquence fondamentale F0, les accélérations et ralentissements du tempo, les changements de registre, les allongements, les pauses (cf. Duez, 1997 ; Fónagy, 1991 ; Touati, 1995, 2003) et, pour citer Léon

(1993, p. 166-169), les autres « avatars de la voix des hommes politiques ». Je me propose ici de préciser les paramètres prosodiques qui permettent de caractériser les différents phonostyles de deux leaders hispanophones. Mon étude est purement descriptive et ne porte pas sur les différences entre les variétés régionales de l'espagnol. Je m'intéresse donc a) à la comparaison des productions de l'ancien président du Vénézuéla, Hugo Chávez (HC) dans deux situations de production (c'est-à-dire, dans deux 'genres' différents) : le discours public 'spontané' et l'interview, et b) à la comparaison des réalisations de HC avec un autre homme politique hispanophone, José L. R. Zapatero (Z) ancien premier ministre espagnol, dans le discours public 'spontané'. Le choix, d'abord intuitif, de ces deux leaders a été confirmé grâce à des tests de perception (en parole naturelle et en parole filtrée) effectués auprès d'auditeurs hispanophones ou non-hispanophones qui les ont classés dans deux catégories différentes (Pérez, 2014). Je terminerai par une brève étude portant sur les imitateurs de ces deux personnalités (cf. Léon 1993, p. 178-179).

2 Méthodologie

2.1 Corpus

Le corpus général est composé de plusieurs enregistrements de discours appartenant à différents genres (interviews, discours à l'ONU ou discours publics pris sur internet) prononcés par 30 hommes et femmes politiques hispanophones d'Amérique Latine et d'Espagne. Ces enregistrements totalisent une quarantaine d'heures. Pour chaque personnalité, une vingtaine de phrases syntaxiquement bien formées et représentatives de la façon dont ils parlent, ont été extraites pour faire l'analyse prosodique. Pour illustrer le style caractéristique de HC et Z, j'utiliserai trois exemples de deux genres différents : interview et discours public 'spontané'. Par ailleurs, j'ai analysé plusieurs sketchs des imitateurs de HC et Z dont j'ai extrait deux exemples. Les analyses acoustiques ont été réalisées avec le programme d'analyse-synthèse 'WinPitch' de P. Martin.

2.2 Modèle intonatif et interprétation

L'interprétation des mesures prosodiques est basée sur le modèle de Ph. Martin *Contraste de Pente Mélodique* et *Structure Prosodique Incrémentale* (Ph. Martin, 1975-2015). Ce modèle propose que les mots 'prosodiques' (définis comme un groupe d'un ou de plusieurs mots avec une seule syllabe accentuée) qui ont une 'relation de dépendance' prosodique entre eux présentent un contraste de pentes mélodique sur leurs dernières syllabes accentuées respectives ; dans un énoncé assertif, un premier niveau de contraste apparaît entre la pente descendante du contour final terminal ('C0') et le contour montant ('C1') le plus proéminent des mots prosodiques précédents. À l'intérieur de ces deux groupes prosodiques il y a un autre niveau possible de contraste entre les mots prosodiques. S'il n'y a pas de relation de dépendance entre les mots prosodiques, les pentes mélodiques sont parallèles comme dans le cas des énumérations. La structure prosodique peut correspondre à la structure syntaxico-discursive (i.e. y être congruente). Mais la structure prosodique est autonome et peut ne pas correspondre du tout à la structure syntaxique, en obéissant par exemple à certains genres de discours, aux phonostyles de chaque locuteur. Par ailleurs la réalisation produite est spécifique à chaque langue, en accord avec la position de l'accent de mot[1].

[1] En espagnol, 80% des mots portent l'accent sur l'avant-dernière syllabe (cf. Quilis 1981). La syllabe accentuée y est alors réalisée par un mouvement mélodique montant ou descendant, tandis que la syllabe finale inaccentuée porte un ton flottant qui va dans le même sens que le mouvement mélodique précédant ou est réalisée par un mouvement de pente contraire pour des raisons stylistiques.

Les différents contours mélodiques observés sont ainsi systématisés dans ce modèle phonologique : **C0** descendant (bas) sur la dernière syllabe accentuée (énoncés assertifs et questions partielles) et éventuellement sur la (ou les) syllabe(s) inaccentuée(s) suivante(s) du mot ; **C1** contour montant, au-dessus du seuil de glissando (voir la formule de glissando dans Rossi 1971) ; **C2** contour descendant non final, au-dessus du seuil de glissando ; **Cn** neutralisé : le contour est légèrement montant ou descendant, avec une durée vocalique plus courte, au-dessous du seuil de glissando, donc la variation mélodique n'est pas perçue.

Les contours complexes : Dans la séquence tonale finale composée d'une syllabe accentuée suivie d'une ou plusieurs syllabes inaccentuées, il peut y avoir une séquence complexe de tons : **C∧** (montant-descendant, au-dessus du seuil de glissando) ou **Cc** (descendant-montant, plat ou légèrement descendant sur la syllabe accentuée et montant sur la syllabe suivante et finale qui est inaccentuée. Pour cette étude, un contour 'Hugo Chávez' **Ch** (contour de nature phonétique semblable au C0 phonologique) a été ajouté. Ce contour spécifique à HC est descendant très bas (chute supérieure à une octave sur la dernière syllabe du mot) à la fin de chaque GI (groupe intonatif composé de un ou plusieurs mots prosodiques) suivi d'une pause et accompagné d'un allongement qui fait pratiquement le double de la syllabe accentuée.

3 Analyse acoustique

3.1 Hugo Chávez en interview

La production d'HC retenue pour les situations d'interview a été diffusée sur la chaîne Univisión en 1998 avant que HC ne soit élu président ; il est interrogé sur ses projets dans le cas où il serait choisi président. L'enregistrement est de bonne qualité, le bruit de fond est d'environ 40 db inférieur à l'amplitude maximale du signal ce qui donne un rapport signal bruit satisfaisant.

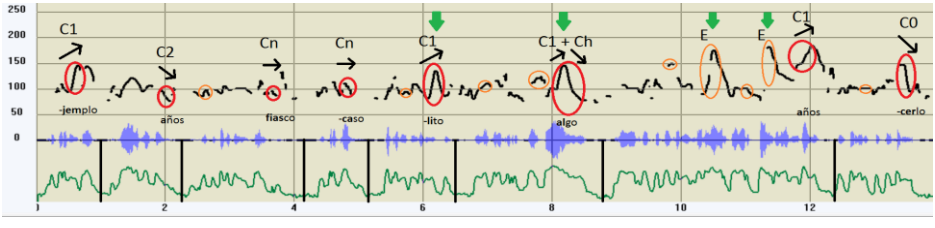


FIGURE 1 : Hugo Chávez en interview

"Si por e**jem**plo (C1) # yo a los dos **a**ños (C2) # re**sul**ta (Cn) que soy un **fias**co (Cn) # un fra**ca**so (Cn) # o co**me**to (Cn) un de**li**to (C∧) # un **he**cho de corrup**ción** (C1) o **al**go (C1+Ch) # que justi**fi**que (C1) mi sa**li**da (C1) del po**der** **an**tes (C1) de los **cin**co **a**ños (C1) # yo esta**ría** dis**pues**to a ha**cer**lo (C0)". [Si par exemple, moi au bout de deux ans, il résulte que suis un fiasco, un échec ou que je commets un délit, un acte de corruption, ou quelque chose qui justifierait ma sortie du pouvoir avant les cinq ans, je serais prêt à le faire]. *Les Contours sont entourés dans la ligne de F0 graduée en Hz (en haut). Pour une question de lisibilité, l'alignement du texte est réalisé seulement sous les Contours pertinents. La variation de l'intensité est en bas (selon une échelle de 50 db). Le signal est au milieu et sa ligne de base coincide avec celle de la F0. Les barres noires verticales indiquent la séparation des groupes intonatifs.*

Dans cet énoncé, HC présente une proposition où il énumère des exemples possibles qui pourraient justifier son départ avant la fin légale de son mandat présidentiel et conclut qu'il pourrait

éventuellement le faire. La construction discursive de cet énoncé est en trois parties, elle est du type : '(1) *si par exemple, (2) tel et tel événement* (= suite de parenthèses)*, (3) je quitte*'.

Pour interpréter la structure prosodique nous avons besoin d'une hiérarchie à plusieurs niveaux : l'un correspondant à la succession de C1 qui contraste avec le C0 et d'autres qui contrastent à un niveau inférieur avec de mouvements mélodiques de moindre ampleur. Au niveau supérieur, le dernier contour C0 (contour de modalité déclarative) qui se trouve sur le dernier mot prosodique '*a ha**cer**lo*' contraste avec les C1 (ou C∧) sur '*si por e**jem**plo*', '*de**li**to*', '*corrup**ción***', '*o **al**go*' et '***a**ños*'. Aux niveaux inférieurs, le C2 associé à la séquence '*yo a los dos **a**ños*', contraste avec les contours C1 sur '*corrup**ción***' et '*o **al**go*'. Plusieurs parenthèses (correspondant à '*tel ou tel événement*') se trouvent au milieu de l'énoncé, chacun de leur contour est réalisé avec Cn ou C1, ce qui est typique dans une énumération, les Cn étant considérés comme des contours parallèles.

Plusieurs accents d'insistance sont réalisés sur des mots sémantiquement importants : '*de**li**to*' (crime), '***al**go*' (quelque chose), '*sa**li**da*' (sortie), '***an**tes*' (avant) ; ils sont indiqués par les flèches verticales qui se trouvent en haut de la figure. Tous ces mots se terminent par un contour C1 suivi d'une syllabe inaccentuée, constituant ainsi des countours complexes C∧. Le septième GI atteint les valeurs de F0 les plus elévées et constitue le 'climax' de l'énoncé.

Il y a bien congruence entre la structure syntaxico-sémantique et la structure prosodique, mais on peut observer un passage qui montre bien l'autonomie possible de la structure prosodique. En effet, la fin du GI '*...o algo*' appartient syntaxiquement aux énumérations précédentes, mais il gouverne aussi le GI suivant '*que justifique*' (dans ces cas ambigus, deux structures prosodiques peuvent être associées à l'énoncé). '*...O algo*' est realisé avec un contour final Ch sur '*-go*' suivi par une pause.

On notera que le registre est assez bas (la F0 moyenne est de 110 Hz) avec une étendue étroite de F0 (50 Hz), alors que les syllabes accentuées de fin d'énoncé arrivent à 180 Hz. Le débit de parole est autour de 6 syll/s et les GI sont constitués de 3 à 20 syllabes (dans cet énoncé).

3.2 Hugo Chávez en discour public 'spontané'

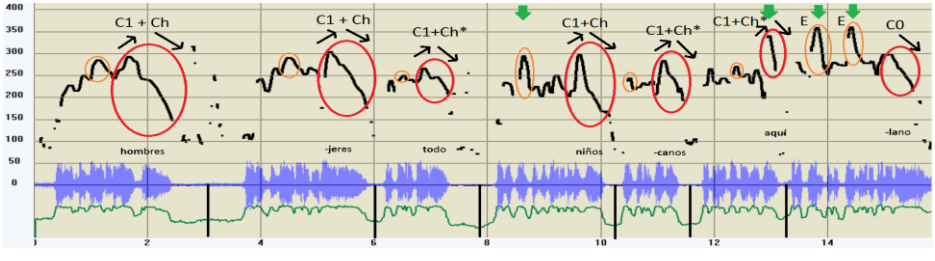


FIGURE 2 : Hugo Chávez en discour public 'spontané'

"A los mi**llo**nes (C∧) de **hom**bres (C1+Ch) # a las mi**llo**nes (C∧) de mu**je**res (C1+Ch) # y sobre **to**do (C1+Ch*) # a los **mu**chos (C∧) mi**llo**nes (C∧) de **ni**ños (C1+Ch) # norteameri**ca**nos (C1+Ch*) # **va**mos (C∧) a man**da**rles (C∧) desde a**quí** (C1+Ch*) # el a**plau**so (C1 ou E) del **pue**blo (C1 ou E) venezo**la**no (C0)". [Aux millions d'hommes, aux millions de femmes, et surtout, à beaucoup de millions d'enfants, nord-américains, on vous envoie d'ici, les applaudissements du peuple vénézuélien].

La réalisation choisie date de 2003. HC prononce ce discours dans le parc "Carabobo" avant un rassemblement contre l'opposition vénézuélienne d'alors où "le principal objectif est de dire non à l'intervention Yankee et à Bush" selon les mots de HC. L'enregistrement est d'assez bonne qualité, et le rapport signal bruit est satisfaisant.

L'énoncé est constitué d'une suite de sept GI séparés par des pauses importantes (de 0.26 à 1.22s). Ils se finissent par un C1 sur la syllabe accentuée du dernier mot prosodique, suivi d'une chute spectaculaire de seize demi-tons sur la dernière syllabe inaccentuée du mot. De plus, cette syllabe inaccentuée est pratiquement deux fois plus longue que la syllabe accentuée précédente (dans la séquence '*a las millones mu**je**res*', l'accent lexical est sur la syllabe '***-je-***' et mesure 0.26s alors que la dernière syllabe '-res' mesure 0.49s). Cette configuration est caractéristique du contour 'Ch', marque phonétique du phonostyle de Chávez. Lorsque ces contours ne descendent relativement pas très bas, ils peuvent être considérés comme des contours de continuation (-*), par exemple dans '*y sobre **to**do**'*, '*norteameri**ca**nos**'*,'*...desde a**quí***'. Du point de vue de la hiérarchie prosodique, au niveau supérieur les contours C1 de continuation sur '*...**hom**bres*', '*...mu**je**res*', '*y sobre **to**do',* '*...**ni**ños*', '*norteameri**ca**nos**'* et '*...desde a**quí***', contrastent avec le C0 final, alors qu'à un niveau inférieur, à l'intérieur de chaque groupe, on retrouve des contrastes de pente, avec des mouvements mélodiques de moindre ampleur, réalisés en C∧, correspondant aux niveaux syntaxiques inférieurs (type 'N de N'). Cette succession de C1+Ch au niveau supérieur contitue une énumeration de groupes prosodiques, ce qui correspond à l'énumeration syntaxique du texte. Il est intéressant de noter que '*norteamericanos*' est considéré, dans la transcription, comme le dernier élément de l'énumeration prosodique avant l'arrivée du climax à la fin de l'énoncé. Mais syntaxiquement on ne sait pas s'il est en relation de dépendance avec tous les syntagmes précédents ou bien s'il s'agit d'un vocatif indépendant.

La F0 moyenne est de 250 Hz et l'étendue du registre va de 150 à 350 Hz (soit plus d'une octave). Le débit de parole est de 4 syll/s et les GI sont formés de 5 à 12 syllabes (dans cet énoncé). Les accents d'insistance se trouvent sur des mots sémantiquement importants, "***mu**chos*" (beaucoup), "*a**quí***" (ici), "*a**plau**so*" (aplaudissement), "***pue**blo*" (peuple), indiqués par les flèches verticales qui se trouvent en haut de la figure.

3.3 Commentaires sur HC dans les deux styles

Dans les deux cas, la ligne de déclinaison n'est pas évidente. En fait, le registre de F0 initial est relativement bas et il monte soudainement jusqu'au point sémantique le plus important (climax) avant de descendre sur le mot final. Ces patrons sont très souvent spécifiques aux discours d'orateurs (cf. Léon 1993, p.166-169).

Les différences trouvées entre les deux situations de productions des discours de HC sont les suivantes : en interview, HC a un patron prosodique 'conversationnel' avec des contrastes de pente de F0 correspondant à la structure syntaxique, un débit de parole plus rapide et une plus petite étendue de la F0. Au contraire, dans le discours public, il fait une énorme chute à la fin de chaque GI, auquel s'ajoute un allongement important de la durée (aux environs de 100%) de la dernière syllabe non-accentuée de chaque GI (cette particularité qui peut être interprétée comme donnant un style 'solennel' se retrouve chez des francophones par exemple le Général de Gaule, cf. Léon 1993 ; Duez 1997). De plus, l'étendue de la F0 est plus importante et le débit de parole est plus lent. Les similarités trouvées dans ces deux phono-genres sont d'un côté les accents d'insistance, et de l'autre le découpage en blocs (GI) ainsi que la forme de la courbe mélodique sur la séquence syllabe accentuée + syllabe inaccentuée (C∧) à la fin de chaque GI, exagérée dans les discours publics

(C1+Ch). Les caractéristiques prosodiques typiques des discours publics de HC peuvent être trouvées de manière régulière tout au long des passages analysés du corpus.

3.4 José Luis Rodriguez Zapatero en discour public 'spontané'

La réalisation choisie pour Z date de 2009. Dans ce discours, il explique devant le Comité Federal du PSOE le plan du gouvernement pour surmonter la crise. L'enregistrement est de bonne qualité et le rapport signal bruit est correct.

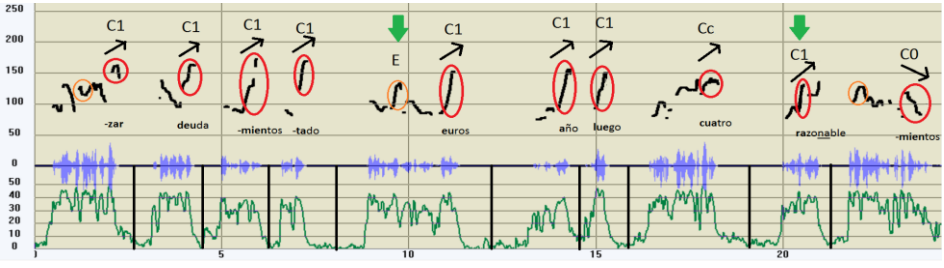


FIGURE 3 : Zapatero en discour public 'spontané'

"Os puedo anun**ciar** (Cn) que vamos a apla**zar** (C1) # el **pa**go(C2) de la **deu**da (C1) # de los ayunta**mien**tos (C1) # con el es**ta**do (C1) # deuda de más de **mil** (C1) quinientos millones de **eu**ros (C1) # que aplazaremos un **a**ño (C1) # y que **lue**go (C1) # de manera razo**na**ble (Cn) durante **cua**tro (Cc) # de manera **ra**zonable (C1) # ten**drán** (C1) que devol**ver**se (Cn) por parte de los ayunta**mien**tos (C0)". [Je peux vous annoncer que nous allons reporter le paiement de la dette, des municipalités, avec l'état, dette de plus de mil cinq cent millions d'euros, que nous reporterons d'un an, et qu'après, de manière raisonnable pendant quatre (ans), de manière raisonnable, devrons être retournés par les municipalités].

Dans les discours publics de Z, au niveau supérieur de la hiérarchie prosodique, les GI sont des petits blocs généralement caractérisés par des contours montants (C1). Ces contours C1 sont dans la grande majorité des cas montants sur l'avant-dernière syllabe (schéma accentuel de l'espagnol) avec une continuation de la montée sur la dernière syllabe non-accentuée. Quelquefois, Z produit des contours complexes Cc comme dans le huitième GI '*de manera razonable durante **cua**tro*'. Il est intéressant de voir qu'occasionnellement, et par exemple dans le neuvième GI *'de manera **ra**zonable*', l'accent de mot n'est pas réalisé à la place attendue et est remplacé par un accent d'insistance sur la première syllabe du mot '***ra**zonable*'. Ce phénomène ressemble à celui décrit comme 'intellectuel' en français (Garde, 1968). Quelle que soit la longueur des GI, il y a un contraste de pente dans les niveaux inférieurs pour indiquer la relation de dépendance entre deux éléments prosodiques. Par exemple, au début de cet énoncé, un zoom sur le deuxième GI montre un C2 sur '*el **pa**go...*' qui contraste avec le C1 final sur '***deu**da*'(voir fig.4).

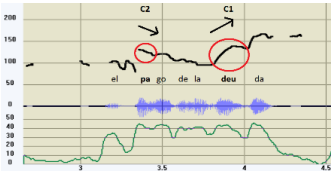


FIGURE 4 : Zoom sur le GI "El **pa**go de la **deu**da" de l'énoncé de la figure 4.

Les réalisations de Z ressemblent en général à une énumération de groupes prosodiques ; les contours (C1) sont parallèles dans un niveau supérieur où ils contrastent tous avec le C0 final. Il n'y a donc pas à ce niveau congruence avec la structure syntaxique. Le registre est assez bas (F0 moyenne = 110Hz) avec une étendue étroite de F0 (40 Hz), sans inclure les pics des syllabes accentuées. Dans cet énoncé aucun allongement important n'a été trouvé. Le débit de parole est de 4 syll/s et les GI sont de 2 à 15 syllabes. Ce patron prosodique 'typique' peut être trouvé de manière régulière tout le long des discours publics 'spontanés' de Z analysés dans ce corpus.

3.5 Imitateurs

Une sélection de quelques imitateurs de ces deux leaders a été faite pour comparer leurs réalisations à celles de HC et Z (ici juste un imitateur pour chacun des deux). Il est intéressant de voir comment ces professionnels de la parole savent caractériser prosodiquement les discours des personnalités.

3.5.1 Imitateur HC (Emilio Lovera)

Chez l'imitateur de HC, nous pouvons bien reconnaître les contours C1+Ch (entourés de cercles sur la figure) typiques de HC ainsi que l'allongement de la dernière syllabe de chaque GI. Ce comédien a bien saisi le phonostyle de HC et l'utilise avec régularité.

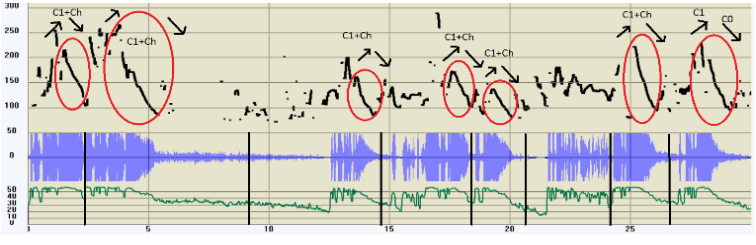


FIGURE 5 : Imitateur de HC, Emilio Lovera

3.5.2 Imitateur Z (Pepe Plaza)

Quant à l'imitateur de Z, il reproduit bien et de manière régulière les contours typiques C1 de Z, ainsi que les GI très petits qui lui sont propres. Puisque le but de ses imitations est de faire rire les spectateurs les réalisations prosodiques sont exagérées et sont répétées sans fin. A titre d'exemple, le dernier contour qui devrait être un contour descendant C0 de modalité déclarative est aussi montant comme tous les autres.

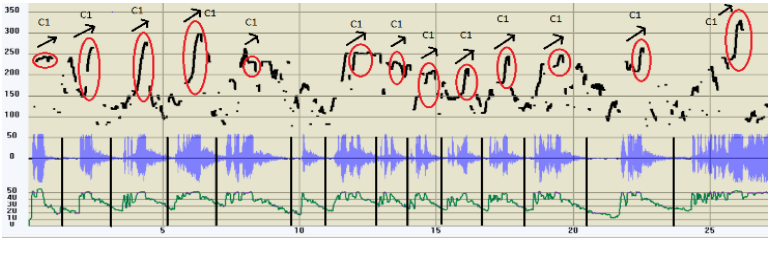


FIGURE 6 : Imitateur de Z. Pepe Plaza

Les imitateurs de HC et Z 'se mettent dans la peau' de ces leaders ; ils savent retrouver et utiliser les traits pertinents du phonostyle de ces personnages et les exagèrent pour faire rire leur public (cf. Carton 1992). Les imitateurs peuvent en fait être considérés comme de véritables 're-synthétiseurs de parole' et leurs réalisations nous servent, dans une sorte d'analyse par synthèse, à confirmer les caractéristiques des phonostyles analysés.

4 Discussion générale et conclusion

Les patrons prosodiques décrits chez HC et Z présentent clairement deux phonostyles différents : les contours typiques de HC sont de type « C1+Ch » alors que celui de Z sont des « C1 » qui continuent vers une valeur plus haute sur la syllabe suivante inaccentuée. Leur phonostyle peut être facilement reconnu et je propose une représentation en figures 7 et 8. D'autre part, les discours de HC et de Z sont généralement coupés en petits blocs ce qui peut être interprété comme une façon de produire les énoncés pour faciliter la compréhension du discours et pour maintenir éveillée l'attention des auditeurs. Il est aussi intéressant de remarquer que cela rappelle le phonostyle de l'ancien président français N. Sarkozy (cf. Martin 2010). HC et Z respectent le patron intonatif général de l'espagnol décrit par les auteurs (Quilis 1999, Sosa 1999, Hualde & Prieto 2015, Martin 2015) mais en exagérant les paramètres prosodiques usuels. De plus, comme les situations de discours (phono-genres) ne sont pas exactement les mêmes et le public différent dans les deux cas, nous pouvons supposer que ces deux phonostyles sont caractéristiques de ces deux politiciens qui ont une histoire et un parcours professionnel différents (HC militaire et Z avocat), sans oublier leur personnalité différente. Je postule que les différences concernant leurs parcours social et politique sont transmises par un phonostyle différent dans leurs discours respectifs. Les résultats de cette analyse phonostylistique des deux leaders politiques hispanophones sont comparables à ce qui a été décrit pour les hommes politiques français (Duez 1997, Touati 1995, 1999, 2003, Léon 1993). A ma connaissance, mon étude est la première portant sur le discours des personnalités politiques hispanophones.

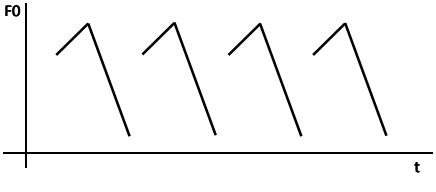


FIGURE 7 : Patron prosodique de HC

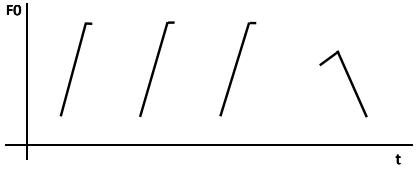


FIGURE 8 : Patron prosodique de Z

Références

CARTON F. (1992). Imitateurs et hommes politiques. (P. Martin, Éd.) Toronto : Mélodie-Toronto.

DUEZ D. (1997). Acoustic markers of political power. *Journal of psycolinguistic research, 26*, 641-654.

FELDHAUSEN I., DELAIS-ROUSSARIE E. (2012). La structuration prosodique et les relations syntaxe/prosodie dans le discours politique . *Journées d'études sur la Parole*, 9-16.

FONAGY I. (1983). *La vive voix. Essais de psycho-phonétique.* Paris: Payot.

GARDE P. (1968). *L'accent*. Paris: Presses Universitaires de France.

HUALDE J., PRIETO P. (2015). *Intonational variation in Spanish : European and American varietes.* (F. &. Prieto, Éd.) Oxford, United kingdom: Oxford U. P.

LEON P. (1993). *Précis de phonostylistique. Parole et expressivité.* Paris : Nathan.

LEON P. (2009). Nouveau regard sur la phonostylistique. *P.U.F. La linguistique, 45*(1), 159-170.

MARTIN P. (1980). *Pour une théorie de l'intonation. L'intonation est-elle une structure congruente à la syntaxe?* (M. Rossi, Éd.) Klincksieck.

MARTIN P. (2009). *Intonation du français.* Paris: Armand Colin.

MARTIN P. (2010). Intonation in Political Speech: Ségolène Royal vs. Nicolas Sarkosy. *Multimodal communication in political speech, 54-64.*

MARTIN P. (2015). *The structure of spoken language. Intonation in Romance.* Cambridge U. P.

PEREZ C. P. (2014). Prosodic realizations of Hispanic politicians. Poster setion presented at *Laboratory approaches to romance phonology*, Aix-en-Provence, 32-34.

QUILIS A. (1999). *Tratado de Fonología y Fonética Española* . Madrid : Gredos.

ROSSI M. (1971). Le seuil de glissando ou le seuil de perception des variations tonales pour la parole. *Phonetica, 23*, 1-33.

SOSA J. M. (1999). *La entonación del Español*. Madrid : Catedra.

TOUATI P. (1995). Pitch range and register in French political speech. *Proc. of the XIIIth International Congress of Phonetic Science, 4*, 244-248.

TOUATI P. (1999). Rhétorique et prosodie des discours politiques. *XIV Skandinavaviska Romantistkongressen*, 1115-1127.

TOUATI P. (2003). Registre et expansion tonals du français. L'usage rhétorique de la voix dans les discours politiques. *Registre et voix sociale*, 59-78.

Etude acoustique et représentation phonologique du suffixe rhotique /ɚ/ en mandarin

Anqi LIU
Université Paris Diderot, CLILLAC-ARP, 5 rue Thomas Mann, 75013 Paris, France
anqi.liu@linguist.univ-paris-diderot.fr

RESUME

Historiquement, le suffixe /ɚ/ est un suffixe diminutif correspondant au mot 儿 (<er> en pinyin) qui signifie "petitesse". Il relève d'une particularité du style plutôt que de la grammaire. Il apparait souvent dans la parole des locuteurs du nord de la Chine. Pour mieux comprendre le phénomène et son comportement phonologique, on présente les résultats d'une étude acoustique qui vérifie les effets de la rhoticité sur les voyelles adjacentes. Sur la base de ces résultats, on propose une représentation gestuelle du suffixe et des processus qui l'impliquent dans le cadre de la phonologie articulatoire (Browman & Goldstein1992).

ABSTRACT

Acoustic study and phonological representation of the rhotic suffix /ɚ/ in mandarin
Historically, the rhotic suffix /ɚ/ in Mandarin is a diminutive suffix corresponding to the word 儿 (er) "smallness". It is considered a marker of style more than of grammar. The rhotic suffix occurs often in the speech of Mandarin speakers from northern China. In order to understand this phenomenon and its phonological behavior, I present the results of an acoustic study, determining the effects of the rhoticity on adjacent vowels. Based on these experimental results I propose a gestural phonological representation of the suffix in the framework of articulatory phonology (Browman & Goldstein1992).

MOTS-CLES : Suffixe rhotique, Mandarin, Analyse acoustique, Phonologie articulatoire
KEYWORDS: Rhotic suffix, Mandarin, Acoustic analysis, Articulatory Phonology

1 Introduction

Le suffixe rhotique est une caractéristique du dialecte de Pékin. Historiquement, le suffixe /ɚ/ correspond au mot 儿(er) qui signifie "petitesse", donc il est considéré comme un suffixe diminutif. Il peut aussi s'ajouter à des adjectifs et des verbes. Pourtant, il n'est pas un suffixe grammatical derivationnel. Il s'agit d'une particularité du style plutôt que de la grammaire. Ce suffixe se combine avec la syllabe à laquelle il s'attache, comme dans les exemples du tableau 1.

sans suffixe	avec suffixe	exemples	traduction
i (y)	(j)əɚ	tɕi/tɕəɚ	‘poulet’
a (u ɤ)	aɚ	pa/paɚ	‘manchon’
ai (əi)	aɚ	$p^ha/p^haɚ$	‘plaque’
an (ən)	aɚ	$p^han/ p^haɚ$	‘assiette’
au (əu)	au^r	tau/tau^r	‘couteau’
aŋ (əŋ, uŋ)	$aŋ^r$	kaŋ/$kaŋ^r$	‘jarre’
in	(j)əɚ	tɕin/tɕəɚ	‘aujourd’hui’

TABLEAU 1 : EXEMPLES DE DUANMU (2007)

L'occurence de ce suffixe est très variable dans la parole. L’objectif de cette étude est de déterminer le comportement phonétique du suffixe basé sur une étude acoustique du contexte où il apparaît, pour ensuite mieux comprendre son comportement phonologique.

Il existe plusieurs analyses phonologiques de ce suffixe et des changements qu’il entraine dans la syllabe à laquelle il s’attache (Lin 1989 ; Duanmu 1990, 2007 ; Wang 1993). L’analyse de Duanmu 2007 est la plus complète et récente. A partir de cette analyse j’effectue une étude acoustique. Basé sur les résultats je propose une représentation gestuelle des mêmes phénomènes, dans le cadre de la phonologie articulatoire (Browman et Goldstein 1992).

Duanmu (2007) propose une règle et deux contraintes phonologiques :

- Règle : Ajouter le suffixe /əɚ/ à la place de la coda. Si le segment de la coda n'est pas compatible avec le suffixe, ce dernier remplace la coda. Sinon /əɚ/ est ajouté directement.
- Contrainte Harmonie de la rime (Rhyme-Harmony) : les segments de la rime sont [retroflex].
- Contrainte d’aperture (Mid) : le degré d’aperture par défaut du noyau vocalique est moyen.

Dans les sections 2, 3, 4 je présente l’étude acoustique sur le suffixe rhotique dans le texte lu, censée vérifier l’analyse phonologique de Duanmu (2007). Je présente les résultats de mon expérience et je les compare aux descriptions de Duanmu. Dans la section 5, je propose une représentation gestuelle préliminaire des données basée sur les généralisations. La conclusion est présentée dans la section 6.

2 Étude acoustique

L’étude acoustique est censée vérifier les données et montrer comment ce suffixe se combine avec le segment précédent. J’adopte la notation de Duanmu 2007, où le symbole /ɤ/ indique la voyelle centrale et /ə/ indique la composante centrale des diphtongues. Le mandarin standard contient 5 voyelles: voyelles fermées /i,y,u/ ; voyelle centrale /ɤ/ ; voyelle ouverte /a/.

Dans une étude acoustique précédente, Huang (2010) a observé pour les monophtongues les changements suivants dans la présence du suffixe rhotique:

- 1) F3 descend pour toutes les voyelles ;
- 2) Pour les voyelles fermées antérieures /i, y/, F1 monte et F2 descend ;
- 3) Pour la voyelle centrale /ɤ/, F1 et F2 montent ;
- 4) Pour la voyelle ouverte /a/, F1 et F2 descendent ;
- 5) Pour la voyelle fermée postérieure /u/, aucun changement important pour F1 et F2.

Huang propose que, du à l'arrondissment de /u/, aucun changement n'est trouvé pour cette voyelle.

Dans mon étude je teste la prédiction que toutes les voyelles vont se centraliser sous l'influence du suffixe rhotique. Je compare les valeurs des trois premiers formants pour déterminer le changement de la qualité vocalique sous l'influence du suffixe rhotique.

2.1 Méthode

Participants. Six locuteurs chinois – trois femmes et trois hommes (entre 25 et 30 ans) – ont été enregistrés. Deux femmes viennent du nord de la Chine et vivent en France depuis 4 ans. A part le mandarin, elles parlent français et anglais comme langues secondes. Trois hommes et la troisième femme viennent du nord de la Chine et vivent en Chine. Parmi eux, un locuteur a eu 5 ans d'expérience aux Etats-unis. A part le mandarin, ils parlent anglais comme langue seconde.

Corpus. Nous avons créé une liste de 78 mots à enregistrer. La moitié (39 mots) contient des monophtongues, des diphtongues et des voyelles orales suivies par une coda nasale, avec les trois tons simples T1, T2 et T4. On exclut T3 qui est un ton sandhi et, par conséquent, sujet à une grande variabilité inter-locuteurs. Les autres 39 mots sont les mêmes avec en plus le suffixe rhotique ajouté. Chaque mot a été enregistré dans une phrase porteuse : Wo214 ba 214 "da55" xie 214 hao214 (j'écris bien "da55"). Les tons du mandarin sont présentés ici par des chiffres : T1 (ton plat) par 55, T2 (ton montant) par 35, T3 (ton descendant-montant) par 214, T4 (ton descendant) par 51. La consonne /t/ a été choisie comme contexte consonantique pour deux raisons. D'une part, l'usage de la même consonne permet de contrôler l'influence de la coarticulation. D'autre part, /t/ est la consonne qui se combine avec la plupart des voyelles du mandarin. Par contre, on a utilisé la consonne /l/ pour la voyelle /y/, parce que cette voyelle se combine avec relativement peu de consonnes. Les phrases ont été présentées imprimées dans un ordre aléatoire. 5 répétitions des phrases ont été enregistrées pour chaque locuteur.

Enregistrements. Les deux enregistrements faits à Paris ont eu lieu dans une chambre sourde, sur un ordinateur, avec le logiciel Praat (à 44100 Hz), via une carte son Edirol et un microphone externe (Audio technica AE4100). Les quatre enregistrements en Chine ont été faits avec le meme équipement dans une chambre à la maison, sans interférence d'appareils électroniques. Les données ont été analysées dans Praat (version 5.3.56) (Boersma and Weenink).

2.2 Analyse acoustique

Une analyse acoustique a été faite par Svantesson (1984) sur la structure formantique des voyelles du mandarin. Dans mon étude, j'ai adopté la même méthode pour avoir une référence des valeurs formantiques. Les trois premiers formants des monophtongues avec et sans suffixe ont été mesurés à la main. Dans le contexte sans suffixe, le début de la voyelle est marqué au début de F1 après le relâchement du /t/. La fin de la voyelle est marquée à la fin de F2 avant le début de friction dans le mot suivant. Dans le contexte avec suffixe, la fin de la voyelle est marquée au début de la chute de F3. Les valeurs des trois premiers formants sont mesurées au milieu de la voyelle segmentée d'après

ces critères.

3 Résultats

3.1 Effets du suffixe rhotique sur les monophtongues

Les formes avec et sans suffixe observées dans nos données sont présentées dans le tableau 2. Pour les voyelles /a, ɤ, u/, aucun changement n'apparaît entre les formes avec et sans suffixe. Les voyelles /i, y/ deviennent plus courtes et un /ə/ apparait dans le signal lorsque le suffixe est présent.

Monophtongue seule	a	ɤ	i	y	u
Monophtongue + Suffixe rhotique	a^r	$ɤ^r$	$jə^r$	$ɥə^r$	u^r

TABLEAU 2 : LA FORME DES MONOPHTONGUES SEULES ET AVEC SUFFIXE RHOTIQUE

Les résultats sont présentés séparément pour les trois femmes (fig. 3A, B) et les trois hommes (fig. 4A, B). Les trois tons sont représentés par trois couleurs différentes : T1 noir, T2 rouge, T4 bleu. ('e' représente /ɤ/ dans la figure). Puisque ces résultats sont encore préliminaires, on ne présente pas d'analyse statistique.

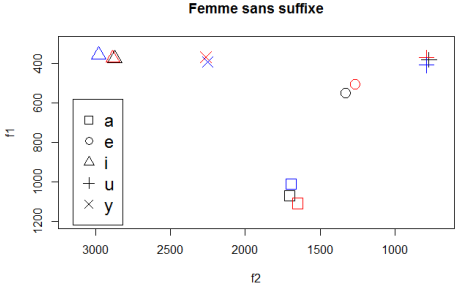


FIGURE 3A : LA MOYENNE DE F1 ET F2 DES TROIS FEMMES POUR LES VOYELLES SEULES

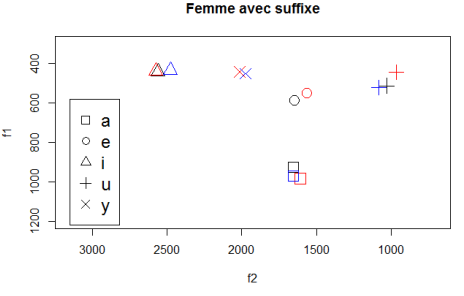


FIGURE 3B : LA MOYENNE DE F1 ET F2 DES TROIS FEMMES POUR LES VOYELLES AVEC SUFFIXE

Les figures 3A et 3B montrent les valeurs moyennes des formants pour les trois femmes, séparément pour chaque ton. La comparaison des deux conditions montre une tendance de centralisation de toutes les voyelles en présence du suffixe. Pour les voyelles fermées /i, y, u/ F1 monte lorsque le suffixe s'ajoute, alors que F1 de /a/ descend. F1 de /ɤ/ monte légèrement. F2 monte pour /u/ et /ɤ/ et descend pour /i/, /y/ et /a/. La dispersion vocalique se centralise quand le suffixe s'ajoute. Au niveau du ton, aucun changement important n'est observé.

Les figures 4A et 4B montrent les valeurs moyennes des formants pour les trois hommes. Comme pour les femmes, la dispersion vocalique tend vers une centralisation lorsque le suffixe s'ajoute. F1 de /i/, /y/, /u/, /ɤ/ a tendance à monter, F1 de /a/ descend. F2 de /u/ et /ɤ/ monte, mais descend pour /i/ et /y/. F2 de la voyelle /a/ a une tendance légère à monter. Au niveau du ton, on n'a pas observé d'effet important.

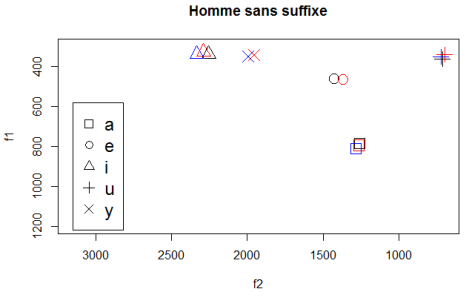


FIGURE 4A : LA MOYENNE DE F1 ET F2 DES TROIS HOMMES POUR LES VOYELLES SEULES

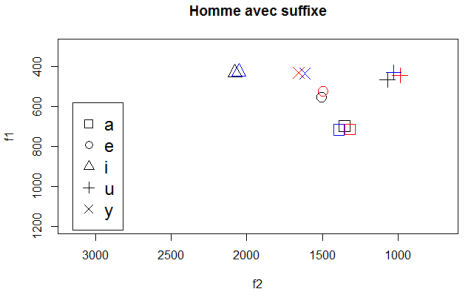


FIGURE 4B : LA MOYENNE DE F1 ET F2 DES TROIS HOMMES POUR LES VOYELLES AVEC SUFFIXE

On observe les mêmes tendances dans les valeurs des formants quand le suffixe rhotique s'ajoute :

– F1 a tendance à descendre pour la voyelle ouverte /a/, qui devient moins ouverte, alors que les voyelles fermées /i, y, u/ ont tendance à devenir moins fermées (F1 monte). La voyelle centrale subit moins de changements.

– F2 des voyelles postérieures /u, ɤ/ a tendance à monter. Les voyelles deviennent moins postérieures. F2 descend pour les voyelles /i, y/, qui deviennent moins antérieures. Pour la voyelle /a/, les hommes ont F2 beaucoup plus bas que les femmes, donc, F2 monte pour les hommes et descend pour les femmes. Mais F2 de la voyelle /a/ a une tendance à se centraliser vers 1500Hz.

Le tableau 5 contient les valeurs moyennes de F3 pour chaque voyelle seule et avec suffixe. On observe que F3 descend pout toutes les voyelles lorsque le suffixe est présent.

Voyelle	a (F/H)	ɤ (F/H)	i (F/H)	u (F/H)	y (F/H)
F3 sans suffixe	2911/2623	3023/2675	3649/3120	2936/2728	2828/2419
F3 avec suffixe	2870/2439	2918/2638	3210/2757	2792/2462	2689/2317

TABLEAU 5 : LA MOYENNE DE F3 POUR LES HOMMES(H) ET LES FEMME (F)AVEC/SANS SUFFIXE

Tous les résultats indiquent que le suffixe rhotique a tendance à centraliser les monophtongues.

3.2 Effets du suffixe rhotique sur les diphtongues

Les diphtongues en mandarin combinent la voyelle centrale /ə/ ou la voyelle ouverte /a/ avec une voyelle fermée. Le mandarin standard contient 4 diphtongues (Duanmu 2007) /ai, əi, au, əu/. Les effets du suffixe rhotique sur les diphtongues sont observés par une analyse qualitative des spectrogrammes dans Praat. Le tableau 6 contient la transcription basée sur les enregistrements.

Diphtongue seule	ai	əi	au	əu
Diphtongue + Suffixe rhotique	a^{r}	ə^{r}	au^{r}	əu^{r}

TABLEAU 6 : La forme des diphtongues seules et avec suffixe rhotique

On a trouvé que /i/ dans la diphtongue disparait acoustiquement lorsque le suffixe est ajouté. En comparant les spectrogrammes de la monophtongue /a/ et de la diphtongue /ai/ avec le suffixe rhotique (FIGURES 8A et 8B), on trouve qu'il n'existe pas de différences entre ces deux figures. Nous confirmons donc que la nouvelle forme avec le suffixe rhotique pour la diphtongue /ai/ est /aʳ/. Il en est de même pour /əi/.

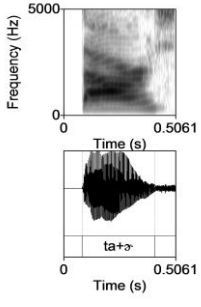
Pour les diphtongues /au/ et /əu/, nous n'avons trouvé aucun changement pour F1 et F2 en présence du suffixe. Cependant, on a remarqué que F3 descend après une courte partie stable. Le seul changement, celui de F3, est causé par l'influence de la rhoticité. Nous pouvons dire que le suffixe rhotique s'ajoute directement aux diphtongues /au/ et /əu/.

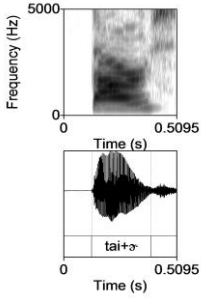
3.3 Voyelle (V) + Nasale (N)

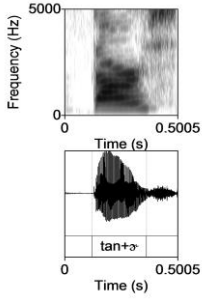
Voyelle(V) + Nasale(N)	an	aŋ	əŋ	uŋ
Voyelle(V) + Nasale(N) +Suffixe rhotique	aʳ	aŋʳ	əŋʳ	uŋʳ

TABLEAU 7 : La forme sans et avec suffixe rhotique pour les cas V+N

Auditivement, il n'y pas de trace de la nasale alvéolaire /n/ quand le suffixe rhotique s'ajoute. Par contre, la nasale vélaire /ŋ/ reste. Pour confirmer la disparition de /n/ alvéolaire, on a comparé le spectrogramme de /tanʳ/ avec celui de /taʳ/ (Figures 8A et 8C). La nasale n'est pas visible dans aucun des spectrogrammes. Pour les cas V+/ŋ/, on a trouvé qu'il existe encore une partie visible de la nasale en présence du suffixe rhotique.







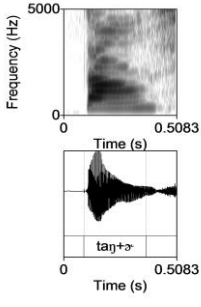


FIGURE8A : ta+ə˞/taʳ/ FIGURE8B : tai+ə˞/taʳ/ FIGURE8C : tan+ə˞/taʳ/ FIGURE8D :taŋ+ə˞/taŋʳ/

L'examen qualitatif du signal montre la présence d'anti-formants pour la coda nasale vélaire aussi bien dans /taŋ/ que dans /taŋʳ/.

4 Discussion

Nous avons comparé les résultats et les observations de notre étude acoustique avec l'analyse phonologique du suffixe rhotique proposée par Duanmu (2007). Les changements correspondent à ceux qu'il décrit. L'analyse spectrale des monophtongues confirme aussi pour la plupart les résultats

de Huang (2010), sauf pour la voyelle /u/. Dans les données examinées ici, toutes les voyelles, y compris /u/, se centralisent sous l'influence du suffixe rhotique.

5 Proposition de représentation gestuelle du suffixe rhotique

Les données des six locuteurs ne sont pas entièrement expliquées par l'analyse phonologique de Duanmu (2007). Notamment, la chute de la coda /n/ apicale, et la retention de la coda /ŋ/ vélaire ne sont pas expliquées dans son analyse. Je propose ici une représentation alternative dans le cadre de la phonologie articulatoire (Browman et Goldstein 1992). Dans l'étude acoustique, on a trouvé que les monophtongues se centralisent sous l'influence du suffixe rhotique. Cependant, les figures 3A et 3B montrent que la distribution des voyelles est toujours dispersée même si le système vocalique s'est beaucoup réduit. Ainsi, les cinq monophtongues restent les mêmes quand le suffixe s'ajoute, sauf pour la distribution de /i, y/. Pour les diphtongues et les cas V+N, on trouve que la coda /i/ et /n/ disparaissent, mais la coda /ŋ/ reste.

Je vais formuler ces généralisations dans le contexte de la phonologie articulatoire. Ce modèle est particulièrement pertinent car il s'appuie sur des observations empiriques. Selon Browman et Goldstein (1992), la parole est composée de gestes articulatoires combinables, les gestes sont à la fois des unités de contraste lexical et des unités physiques d'action articulatoire. Ici, on représente le suffixe rhotique [ʳ] comme étant composé de deux gestes articulatoires. L'un est le relèvement de la pointe de la langue vers le palais, l'autre est la rétraction du dos de la langue vers le bas et l'arrière. Les généralisations sont résumées dans le Tableau 9A. Il en résulte 3 cas: 1) Rajout de /ə/ dans le contexte /i/ et /y/ ; 2) Rajout du suffixe rhotique ; 3) Chute de coda /n/ et /i/.

	Sans suffixe	Avec suffixe
1. Rajout de /ə/ dans le contexte /i/ et /y/	i, y	jəʳ , ɥəʳ
2. Rajout du suffixe rhotique	a, u, ɤ	aʳ, uʳ, ɤʳ
	au, əu	auʳ, əuʳ
	aŋ, əŋ, uŋ	aŋʳ, əŋʳ, uŋʳ
3. Chute de coda /n/ et /i/	an, ən	aʳ, əʳ
	ai, əi	aʳ, əʳ

TABLEAU 9A : GENERALISATIONS BASEES SUR LES RESULTATS DE L'ETUDE ACOUSTIQUE.

Exemple sans suffixe	Exemple avec suffixe	Traduction
ti	tjəʳ	sol
ta	taʳ	grand
tau	tauʳ	couteau
taŋ	taŋʳ	devoir
tan	taʳ	drap
tai	taʳ	sachet

TABLEAU 9B : EXEMPLES AVEC ET SANS SUFFIXE.

Dans ce qui suit, je vais préciser les gestes articulatoires qui caractérisent les voyelles et les nasales des codas pour comprendre comment les changements observés découlent de leurs interactions.

5.1 Rajout de /ə/ dans le contexte /i/ et /y/

La voyelle /i/ est composée de deux gestes : l'un est l'étirement des lèvres, l'autre est l'avancement de la langue. /i/ et /ʳ/ imposent des contraintes linguales contradictoires. La langue s'avance pour /i/ et se lève vers le palais pour la rhoticité, alors que le dos de la langue se retire. Je propose que /ə/ émerge comme une transition entre les deux premiers gestes et la rétraction du dos de la langue. Les

gestes articulatoires pour la production de la voyelle /i/ sont conservés. Dans ce cas, /i/ devient la semi-voyelle /j/, qui a les mêmes gestes articulatoires que /i/, mais racourcis.

/y/ est composé de deux gestes, l'arrondissment des lèvres et l'avancement de la langue. /y/ et /ʳ/ imposent aussi des demandes linguales contradictoires. Comme dans le cas de /i/, la production de /y/ se racourcit et devient la semi-voyelle /ɥ/.

5.2 Rajout du suffixe rhotique

Les trois cas /aŋ/, /əŋ/, /uŋ/ contiennent la coda nasale vélaire /ŋ/ qui est composée de deux gestes : la rétraction du dos de la langue et l'abaissement du vélum. /ŋ/ partage le premier geste avec le suffixe rhotique. Le geste apical du suffixe peut s'ajouter au geste dorsal et à la baisse du vélum. Ainsi, rien d'autre ne change quand on ajoute le suffixe rhotique.

Les cas /au/, /əu/ et /u/ finissent par /u/ qui est composé de deux gestes, un geste labial d'arrondissement des lèvres et un geste de rétraction du dos de la langue. /u/ partage un geste avec le suffixe rhotique, c'est-à-dire, la rétraction du dos de la langue. Comme dans le cas de la nasale vélaire /ŋ/, le geste apical de la rhoticité s'ajoute sans autres changements.

/a/ et /ɤ/ n'ont pas de geste de retraction ou d'avancement de la langue. La cohésion entre ces voyelles et le suffixe rhotique est donc facile à réaliser. Ainsi, le suffixe va être ajouté directement à ces monophtongues.

5.3 Chute de coda /n/ et /i/

La coda /n/ est composée de deux gestes : l'abaissement du vélum et l'avancement de la pointe de la langue vers les alvéoles. La nasale ne partage aucun geste articulatoire avec le suffixe rhotique. D'une part, la pointe de la langue doit avancer vers les alvéoles, d'autre part, le vélum descend pour la nasalité. La cible du geste apical de /n/ est plus antérieure que celle de la rhotique. Ainsi, comme les deux cibles sont incompatibles, seule celle qui correspond au suffixe rhotique est gardée. Les deux gestes rhotiques sont réalisés dans le contexte du suffixe. L'abaissement du vélum est aussi réduit par la rétraction dorsale rhotique. Il est ainsi prédit que la nasalité sera sensiblement réduite.

Les diphtongues /əi/ et /ai/ qui finissent par /i/ ne partagent pas de gestes articulatoires avec le suffixe rhotique. En général, /i/ de la diphtongue devient une semi-voyelle /j/. Mais à cause du geste articulatoire de /a/ et /ə/, la pointe de la langue doit d'abord rester dans une position centrale pour /a/ ensuite elle doit avancer pour /i/, enfin elle doit se relever vers le palais. À cause de sa position et du geste articulatoire qui s'oppose à la rhoticité, la pointe de la langue n'atteint pas la cible de /i/ quand on ajoute le suffixe rhotique. Donc, /i/ disparaît complètement : /tai/ > /taʳ/ 'sachet'.

6 Conclusion

L'analyse de Duanmu (2007) nécessite des contraintes et règles spécifiques au phénomène étudié pour expliquer les données. Une représentation gestuelle comme celle proposée ici permet de prédire les détails observés sans recourir à des règles ou contraintes arbitraires. Elle explique, notamment, la chute de la coda /n/ apicale et la rétention de la coda /ŋ/ vélaire. Les effets de la rhoticité découlent des interactions entre les composantes gestuelles.

Références

BELL-BERTI F. (1993). Understanding velic motor control: Studies of segmental context. *Phonetics and Phonology Vol5: Nasals, nasalization, and the velum*, 63-85.

BROWMAN, C.P., GOLDSTEIN L. (1989). Articulatory Gestures as Phonological Units. *Phonology 6*, 201-251.

BROWMAN, C.P., GOLDSTEIN L. (1990). Gestural Specification Using Dynamically-defined Articulatory Structures. *Haskins Laboratories Status Report on Speech Research*, *SR-103/104*, 95-110.

BROWMAN, C.P., GOLDSTEIN L. (1992). Articulatory Phonology: An Overview. *Haskins Laboratories Status Report on Speech Research 1992, SR-111/112*, 23-42.

CHAO Y. (1968). A Grammar of Spoken Chinese. *University of California Press.* Berkeley.

SAN D. (1990). A Formal Study of Syllable, Tone, Stress and Domain in Chinese Languages. *Doctoral dissertation*. MIT, Cambridge, Mass.

SAN D. (2007). *The Phonology of Standard Chinese (2nd Edition).* Oxford.

HUANG T. (2010). Er-suffixation in Chinese monophthongs: phonological analysis and phonetic data. *Proceedings of the 22nd North American Conference on Chinese Linguistics (NACCL-22) & the 18th International Conference on Chinese Linguistics (IACL-18) 2010. Vol 1*, 331-344.

OLIVE J.P., GREENWOOD A. & COLEMAN J.（1993). *Acoustics of American English Speech: A Dynamic Approach*. Springer-Verlag, New York.

LIN Y.H. (1989). Autosegmental treatment of segmental processes in Chinese phonology. *Doctoral dissertation*. Lieu: University of Texas, Austin.

MAMCHON A., FOULKES P., TOLLFREE L. (1994). Gestural representation and Lexical Phonology. *Phonology 11*, 277-316.

SVANTESSON J-O. (1984). Vowels and diphthongs in Standard Chinese. *Working Papers 27*. 209-335. Lieu : Lund University, Dept. of Linguistics.

WANG J.Z. (1993). The Geometry of Segmental Features in Beijing Mandarin. *Ph.D. dissertation*, University of Delaware, Newark.

WANG L., HE N. (1985). Beijinghua er-huayun de tingbian shiyan he shengxue fenxi. *Beijing Yuyin Shiyanlu*, 27–72.

Étude de la contribution acoustique de la structure formantique à l'identification du ton chuchoté

Xuelu Zhang[1], Rudolph Sock[1, 2]

(1) U.R.1339 LILPA/Equipe de Recherche Parole et Cognition & Institut de Phonétique de Strasbourg (IPS), Université de Strasbourg (UNISTRA), 22, rue Descartes, 67084 Strasbourg, France

(2) Language, Information and Communication Laboratory (LICOLAB), Faculté des Lettres, Université Pavla Jozefa Safarika, Šrobárova 2, 040 59 Košice, Slovaquie

zhang_xuelu@yahoo.fr

RESUME

Cette étude examine la contribution de la structure formantique du segment vocalique à l'identification du ton que ce segment porte, et cela en voix chuchotée. Le mandarin a été choisi en tant que langue cible parce que les traits tonals (*tone features*) en mandarin s'appuient acoustiquement sur deux dimensions : le registre et le contour. Nous supposons qu'en l'absence de F0, la structure formantique subirait néanmoins une modification, en fonction du ton et fournirait des indices acoustiques des traits tonals à l'auditeur. Nous nous intéressons aux rapports entre les deux dimensions de traits tonals et à la modification de la structure formantique. À travers l'analyse des données acoustiques issues de 13 sujets locutrices, nous avons observé une divergence d'importance dans les intervalles F2-F3 et F3-F4, en fonction du ton. Cette divergence semble liée aux contrastes tonals en registre et non aux contours mélodiques. Cette distinction semble dépendre d'ailleurs de la nature de voyelle.

ABSTRACT

A study of the acoustic contribution of formant structure to tone identification in whispered speech

This study explores the contribution of vowel formant structure to tone identification, particularly in whispered speech. Mandarin was chosen for this study as the target language, as its tone features are acoustically built up on two dimensions : register and contour. We posit that in case of absence of F0, the formant structure would always undergo modification associated to tone, and hence provide acoustical tone cues to the listener. Also, we are interested in relations between the two dimensions of tone features and modifications of formant structures. By analysing the acoustic data collected from 13 female speakers, we found remarkable differences between F2-F3 and F3-F4 values, with regards to a specific tone. Such differences seem to be linked to tone register contrasts and not to melodic contours. Furthermore, this divergence in formant pattern intervals seems to depend on the nature of the vowel.

MOTS-CLES : Identification du ton, mandarin, traits tonals, structure formantique, sensation psychoacoustique de la hauteur de la voix, voix chuchotée.

KEYWORDS : Tone identification, Mandarin, tone features, formant structure, psycho-acoustical sensation of pitch, whispered voice.

1 Introduction

Le ton lexical est un phénomène bien étudié dans le domaine de la phonologie. En tant que composante de la structure lexicale, le ton permet de réaliser une distinction au niveau sémantique à travers des patterns figés de modification mélodique.

Ce phénomène suprasegmental varie d'une langue à l'autre. Il est pourtant possible de décomposer tous types de tons en traits tonals (*tone features*). Selon le pattern de modification du pitch, un système tonal pouvait être catégorisé en tant que système de registre (*Level-Pitch Register System*), système de contour (*Gliding-Pitch Contour System*) ou système qui combine ces deux premiers (*Register-Contour Combination*). Depuis le développement de la phonologie autosegmentale, la notion du système de contour est remise en question : on considère que le contour tonal peut être subdivisé davantage. Goldsmith considérait le contour tonal comme une séquence de tons de registre, et l'a décrit en employant des traits binaires. Jusqu'aujourd'hui, beaucoup d'études sur les tons dans des langues asiatiques emploient cette description à traits binaires, mais aussi un système de traits tonals à quatre registres que Yip a proposé en 1980. Dans les études sur le mandarin et sur d'autres variantes du chinois, le système tonal de Chao Y.R. à cinq partitions (soit cinq registres) est toujours d'actualité. Dans toutes ces descriptions, le registre reste la base des traits tonals : le ton est défini par des registres cibles à atteindre.

Cependant, d'autres phonologues insistent sur une description des tons à contours avec la direction du mouvement du pitch en tant que trait tonal. Yip (2001) a révisé le système de traits tonals à registres tout en prenant en compte la réalisation phonétique de différents types de tons, et elle est arrivée à la conclusion que la description des tons à contours, selon un tel système, n'est pas satisfaisante. Phonétiquement, un ton à contour consiste plutôt en une cible à atteindre et un éloignement de la cible, alors qu'un ton à registre se centre sur les cibles et non sur le mouvement.

Nous nous intéressons donc à cette problématique de représentation des traits tonals, et nous considérons qu'il serait probablement nécessaire de repenser la définition des traits tonals suivant l'analyse des réalisations phonétiques du ton. Dans cette étude, nous avons choisi le mandarin en tant que langue cible, parce que dans cette langue monosyllabique, le ton se réalise dans le même cadre temporel que la syllabe, ce qui facilite l'observation et l'analyse. De plus, le système tonal du mandarin présente une combinaison de tons à registres et de tons à contours, et étant donné ce fait complexe, il connaît effectivement des désaccords dans sa description.

Le mandarin possède quatre tons lexicaux. Les études phonétiques et phonologiques existantes ont montré que ces tons se produisent dans deux partitions de toute l'étendue du pitch tonal (*pitch range*). Le 1er ton et le 4ème ton se réalisent, durant toute syllabe, dans une partition plus haute alors que le 2ème et le 3ème ton se réalisent dans une partition plus basse. Cette division est décrite comme [+/-Upper] (+/-U). Les registres cibles des deux tons dans chaque partition sont en contraste,

et leurs contrastes peuvent être transcrits par [high/low] [1] (h/l). Les quatre tons du mandarin peuvent être transcrits comme suit :

- Ton 1(T1) : [+U, h]
- Ton 2(T2) : [-U, h]
- Ton 3(T3) : [-U, l]
- Ton 4(T4) : [+U, l]

L'analyse dans cette recherche reposera principalement sur cette description des tons du mandarin. Et nous tenterons de remettre ces traits tonals en question en l'absence de F0 (le paramètre acoustique principal de la perception du ton). Autrement-dit, le but de cette recherche est de trouver des indices acoustiques des traits tonals, lorsque ces derniers ne peuvent plus se réaliser à travers F0. Afin d'observer des stratégies de compensation spontanée déployées par le locuteur, nous avons exigé des productions en voix chuchotée, lors du collectage des données acoustiques. Il s'avère que lorsque des natifs qui pratiquent une langue tonale communiquent en chuchotant, ils arrivent à se comprendre avec beaucoup d'aisance, malgré la perte d'informations tonales provoquée par l'absence de F0. Ils doivent donc avoir trouvé un autre moyen de coder-décoder les informations relatives aux tons. Et cela se ferait, d'après nous, au niveau articulatori-acoustique grâce à un processus de réafférences sensorielles.

2 Compensation de la perte de F0 en voix chuchotée

La littérature phonétique, dans le domaine du discours chuchoté, en liaison avec les tons, est assez limitée et relativement récente, même si elle est pourtant fascinante. Les chercheurs qui ont étudié la prosodie dans le discours chuchoté ont découvert, en examinant divers paramètres acoustiques, des indices liés aux tons. Selon Segerbäck (1965), l'activité vibratoire des plis vocaux ne serait même pas indispensable pour toute intercompréhension dans une langue tonale. L'auditeur sinophone utilise des indices secondaires (étant donné que F0 fournit des indices primaires) de manière flexible dans le jugement du ton (Liu et al., 2004). La durée du segment vocalique, par exemple, varie en fonction du ton et cette variation pourrait être significative pour la perception du ton (Blicher et al., 1990). Un autre exemple est le contour de l'intensité qui est corrélé avec F0. Il fournirait également des informations tonales (Whalen et al., 1992 ; Chang et al., 2007).

Nous portons un intérêt particulier aux rapports entre la structure formantique du segment vocalique et le ton que ce dernier porte. Malgré le fait que la structure formantique n'influence pas de manière systématique le jugement du ton perçu (Tseng et al., 1986), une modification du spectre existe réellement lorsque le ton varie. Cela implique une modification de la forme du conduit vocal au niveau supraglottique qui, bien-entendu, peut être issue du mouvement vertical du larynx en voix modale. Cependant, l'amplitude des mouvements du larynx est remarquablement réduite lorsqu'on chuchote. Ainsi, l'emploi de la voix chuchotée peut nous fournir de bonnes conditions pour l'observation des gestes liés au ton. Nous supposons que si la structure formantique connaissait une modification en fonction du ton, même en voix chuchotée, cela apporterait davantage d'arguments à l'existence de gestes articulatoires au niveau supraglottique spécifiques, corrélés au ton, ou bien, aux traits tonals.

De plus, du point de vue psychoacoustique, la structure formantique a le potentiel de transmettre des

[1] Si on considère le contour du pitch comme trait principal ici, on peut le décrire par [fall/rise] (Wang, 1967, cité par Zhang, 2014).

informations tonales par la reproduction d'une «impression» musicale. La hauteur du son de la parole que l'humain perçoit est le résultat d'un traitement du son complexe dans le système auditif périphérique et central. On sait que lorsqu'un locuteur emploie une voix modale, le signal de la parole comporte des propriétés, certes hautement complexes, mais périodiques (y compris sinusoïdales), et la hauteur du son perçue chez l'auditeur dépend principalement (mais pas exclusivement) de la réponse du nerf auditif à la stimulation de F0. Et lorsque le locuteur chuchote, le signal de la parole ne contient plus ces propriétés acoustiques périodiques (dont les composantes sinusoïdales), mais plutôt un bruit de bande, modulé en amplitude (de manière irrégulière) et en fréquence. La perception d'un tel bruit chez l'auditeur ressemble à celle d'un son sinusoïdal, à condition que la bande soit assez étroite. Et pourtant, à cause de l'absence de F0, il n'y a plus de réponse du nerf auditif corrélée directement à la hauteur musicale. La hauteur perçue dans ce cas dépend des réponses de plusieurs fibres nerveuses au stimulus acoustique. Plus précisément, les impulsions sur les fibres nerveuses répondent aux fréquences centrales des bandes du bruit. Ces premières diffèrent en durée du cycle. Ainsi dans le système auditif, *«...quand on module en fréquence des bruits de ce type (bruit de bande), on fait varier leur fréquence limite et cette modification peut être perçue.»* (Zwicker et al., 1981). Cette conclusion a été soutenue par des recherches ultérieures sur la contribution du TFS (*temporal fine structure*) à la perception du pitch (à titre d'exemple, voir Kong et al., 2006).

Ce bruit de bande que la voix chuchotée produit est issu de l'effet de filtrage que le conduit vocal a sur le bruit blanc produit au niveau du larynx. Par rapport à la voix modale, le spectre de la voix chuchotée est différent selon les points suivants : 1) les valeurs formantiques sont plus élevées ; 2) l'énergie se diffuse davantage à des hautes fréquences ; 3) l'intensité acoustique est d'environ 20 dB au-dessous de celle de la voix modale. Ces différences impliquent que la redistribution de l'énergie sonore dans le spectre est à prendre en compte dans l'analyse d'une voix chuchotée. Et on se demande ainsi, quels sont les formants les plus affectés par le ton. Sur cette dernière problématique, les études existantes se contredisent. Higashikawa et al. (1999) supposent que la perception du ton chuchoté est influencée par une variation simultanée du F1 et du F2, alors que Meyer-Eppler (1957) et Fónagy (1969) considéraient que le déplacement des formants à des hautes fréquences vers le haut, tels le F3 et le F4, indique une courbe mélodique montante. Leurs données avaient montré, de manière générale, que le contour mélodique était le trait tonal le plus associé à la modification de la structure formantique.

3 Réflexions et hypothèses

Comme il est mentionné ci-dessus, nous portons un intérêt particulier aux rapports entre la modification de la structure formantique et les traits tonals. Cet intérêt est dû principalement au fonctionnement du système auditif dans la perception du bruit de bande et à la distribution de l'énergie sonore dans le spectre de la voix chuchotée. Vu que la hauteur mélodique du bruit de bande ne se réalise pas à travers une seule bande formantique mais plusieurs, l'un des objectifs de cette étude est de mettre au jour les formants qui seraient les plus concernés par un ton chuchoté. L'autre objectif est de mettre en lumière le trait tonal/les traits tonals suivant lequel/lesquels s'oriente la modification de la structure formantique.

Concernant les formants les plus liés au ton, et malgré la position de Higashikawa et al. (1999), nous sommes assez en ligne avec les conclusions de Meyer-Eppler et celles de Fónagy, pour les raisons suivantes : 1) Dans le spectre du signal acoustique du chuchotement, nous avons observé plus d'énergie présente au niveau du F3, du F4 et du F5, alors que peu d'énergie se présente au niveau du F1 ; 2) la variation du F1 et celle du F2 subissent des contraintes que les frontières des

phonèmes vocaliques imposent au niveau articulatoire, ainsi ces formants peuvent être peu efficaces pour indiquer les contrastes tonals, surtout lorsqu'il s'agit d'une langue tonale à vocalisme riche. Etant donné que le « pitch fantôme » de la voix chuchotée, qui n'existe pas dans le spectre mais est reconstitué dans le système perceptif, semble être lié à la modulation en fréquence dans le bruit, nous supposons que la hauteur et la variation de ce « pitch fantôme » se réaliseraient à travers les intervalles entre les fréquences centrales des formants voisins, probablement celui entre les formants à hautes fréquences.

En ce qui concerne le(s) trait(s) tonal(s) qui se présente(nt) en voix chuchotée, nous supposons que la modification de la structure formantique est capable de réaliser, non seulement le contraste en contour, mais aussi celui en registre. Une langue tonale dont les tons sont en contraste uniquement au niveau du registre n'empêche pas les utilisateurs de cette langue de chuchoter ces tons, alors les informations des registres ont la possibilité de se diffuser à travers la structure formantique[2].

4 Protocole expérimental

Les résultats de cette étude s'appuient sur l'analyse de données acoustiques recueillies dans le cadre d'une expérience de production orale lancée en mars 2013 à Beijing en Chine. 13 locutrices d'origine chinoise, âgées entre 18 et 21 ans, ont pris part à l'expérience. Elles maîtrisent parfaitement le mandarin. Pour l'expérience, nous avons établi un corpus contenant 12 phrases en mandarin, chacune portant respectivement les syllabes /pa//pi//pu/, avec l'un des quatre tons dans le même contexte phonétique, soit /pV_ts‿/. Aucun sandhi tonal ne s'est produit dans cette expérience.

Chaque sujet locutrice a produit 10 répétitions de ce corpus, respectivement en voix normale et en voix chuchotée. L'enregistrement des voix des locutrices a été réalisé dans un environnement silencieux, avec un enregistreur Marantz Professional© PMD661 et un microphone Sennheiser e845 S®. Le recueil des données a été réalisé sur PRAAT et le traitement statistique des données a été réalisé sur Excel 2010 et sur GraphPad Prism 5.

5 Résultats expérimentaux

5.1 Analyse des fréquences centrales des bandes formantiques

Afin de suivre la variation spectrale à l'intérieur du segment vocalique, nous avons pris les valeurs des fréquences centrales des bandes formantiques aux points 0%, 20%, 40%, 60%, 80% et 100% de la durée totale du segment et ainsi défini les courbes de variation spectrale en temps normalisé. Des analyses ANOVA à deux facteurs ont été conduites en considérant la moyenne des 10 répétitions de chaque sujet locutrice comme une répétition non appariée, et comme deux facteurs analysés, le *ton* et le *temps normalisé*. Pour les analyses ANOVA, nous donnons les données de variance avec le *ratio F* correspondant à la variabilité entre les sujets et la valeur de *p*. Seuls les résultats avec une probabilité de moins de 5% ($p<0,05$) sont considérés comme significatifs.

Pour une raison de commodité, nous prenons ici la voyelle /a/ chuchotée à titre illustratif (les

[2] Dans la littérature (ex. Whalen et al., 1992 ; Blicher et al., 1990), la divergence d'autres paramètres acoustiques, tels que l'intensité et la durée tonale, nous semble associée davantage au contraste en contour tonal et moins au contraste en registre.

résultats de l'analyse de /i/ et de /u/ sont proches de ceux de /a/, quel que soit le sujet). En outre, nous renvoyons le lecteur à une de nos études connexes qui présentente des données pour des productions en voix normale (Zhang et al., 2015). Cette étude montre des variations des paramètres acoustiques, en fonction du ton, aussi bien en voix normale qu'en voix chuchotée.

La figure 1 présente les courbes de variation des premiers quatre formants de /a/ aux quatre tons du mandarin. En observant les courbes de variation spectrale en fonction du ton sur chaque graphique, nous concluons que le ton a effectivement tendance à influencer la structure formantique. Pourtant les rapports entre les traits tonals et la variation formantique ne sont pas linéaires. Le trait [+/U], qui sépare T1 et T4 de T2 et de T3, semble avoir fait augmenter les valeurs formantiques du F2, F3 et F4. Cette augmentation est relativement faible par rapport à la valeur absolue du formant en question, surtout par rapport à celles des formants à hautes fréquences, et cela s'est effectivement montré non-significatif ($F(1,21)=0,0015$, $p=0,9691$ pour T1 et T4, et $F(1,21)=0,2316$, $p=0,2316$ pour T2 et T3). Le trait [h/l], qui sépare T1 de T4 et T2 de T3, semble avoir légèrement plus d'effet en séparant T2 de T3, et moins d'effet en séparant T1 de T4. D'ailleurs il est difficile de préciser si l'effet du contraste [h/l] sur la variation formantique se manifeste tout au long du segment ou seulement à la fin du segment. Le résultat des analyses statistiques a montré que cet effet n'est pas significatif ($F(1,20)=0,5120$, $p=0,4758$ pour T 1 et T4, et $F(1,22)=0,9885$, $p=0, 3221$ pour T2 et T3).

L'observation de la variation des formants ne nous révèle pas de résultats significatifs relatifs aux rapports entre les traits tonals et les valeurs absolues des formants. Nous allons donc nous tourner vers une observation des intervalles entre les fréquences centrales des formants.

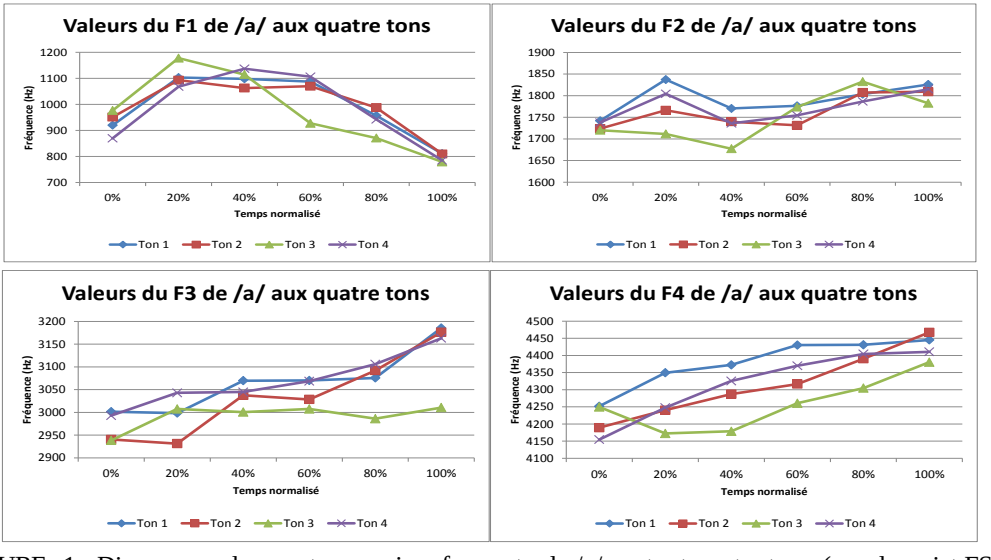


FIGURE 1 : Divergence des quatre premiers formants de /a/ portant quatre tons (cas du sujet FSY).

5.2 Analyse des intervalles entre les fréquences centrales des formants

Dans l'objectif de quantifier l'impression psychoacoustique que la structure formantique pourrait laisser au système auditif, l'analyse des intervalles entre les fréquences centrales des formants repose sur l'emploi de l'échelle *mel*. La conversion de Hertz en *mel* se fait avec la formule bien connue suivante :

$$m = 2595 * log_{10}\left(1 + \frac{f}{700}\right)$$

À titre illustratif des résultats de cette analyse, nous prenons encore le cas de la voyelle /a/. La figure 2 illustre la variation des intervalles entre les fréquences centrales des bandes formantiques voisines de cette voyelle, en fonction du ton.

En observant les courbes sur chaque graphique, nous constatons que les courbes correspondant aux différents tons se distinguent davantage en largeur de l'intervalle (mesuré en *mel*) et peu en contour. Les différences entre elles sont maximales au milieu du segment vocalique. De plus, en comparant les courbes des trois graphiques, nous avons observé que les intervalles F1-F2 et F2-F3 de /a/ aux tons au trait [+U] sont plus importants que ceux aux tons au trait [-U] (sauf pour ce qui concerne la similarité entre l'intervalle F2-F3 de T2 et celui de T4). Cependant, l'intervalle F3-F4 entre les tons est à l'inverse de cet ordre ; les intervalles F1-F2 et F2-F3 de /a/ aux tons au trait [+U] sont, cette fois-ci, moins importants que ceux aux tons au trait [-U]. Le trait [h/l] semble avoir fait séparer davantage les courbes, et selon ces graphiques cette séparation se manifeste davantage au milieu du segment. Une remarque importante s'impose ici. Par commodité, les résultats de /i/ et de /u/ ne sont pas donnés ici. Toutefois, signalons que l'effet du trait [+/-U] sur les intervalles F3-F4 de ces deux voyelles est différent de celui sur /a/ : T1 et T4 provoquent des intervalles plus importants entre le F3 et le F4, par rapport à T2 et T3.

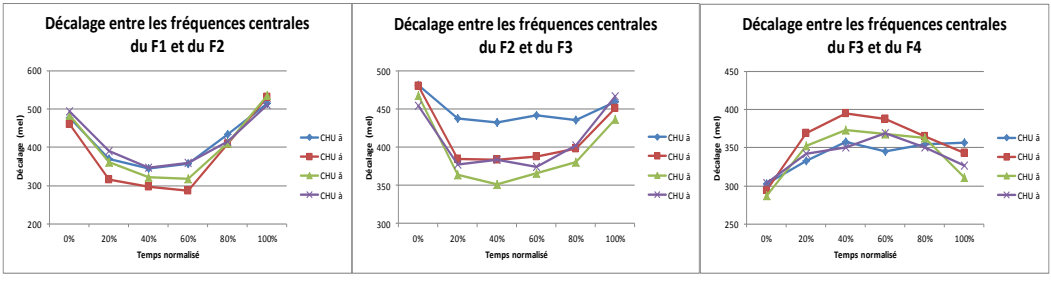


FIGURE 2 : Les intervalles entre les fréquences centrales des bandes formantiques voisines de /a/ en fonction du ton (cas du sujet ZBJ) : /ā/ =/a/+T1, / á/ =/a/+T2, /ǎ/ =/a/+T3, / à/ =/a/+T4.

Pour connaître davantage l'influence du ton sur la structure formantique, nous avons effectué une analyse statistique de tous ces intervalles entre les formants en fonction du ton. Le résultat des analyses statistiques (p=*ns*) montre que la variation du ton ne provoque pas systématiquement de différences significatives sur les intervalles entre les formants, et pourtant ces intervalles ont des sensibilités différentes à la variation tonale. Le tableau 1 illustre le nombre de sujets (sur la base de 13 sujets) qui effectuent des différences significatives lorsque le ton varie de l'un (marqué en rang, transcrit en traits tonals) à l'autre (marqué en colonne, transcrit en traits tonals).

En comparant les cas des trois voyelles dans ce tableau, nous avons remarqué, chez les 13 sujets, que la variation tonale provoque plus de différences significatives ($p<0,05$), dans l'intervalle entre les formants dans la réalisation de /a/, et moins dans la réalisation de /i/ (par ex., des différences significatives ont été trouvées dans l'intervalle F3-F4 chez 7 sujets sur 13, lorsqu'elles réalisent /a/, respectivement à T1 et à T3, alors 4 sujets sur 13 ont fait des différences significatives dans l'intervalle F3-F4 pour distinguer /i/ à T1 de /i/ à T3). En comparant les intervalles entre différentes paires de formants, nous avons remarqué que l'intervalle F3-F4 est le plus sensible au contraste tonal, alors que l'intervalle F1-F2 est le moins sensible (ex. le contraste entre T1 et T3 sur /u/

suscite des différences significatives dans l'intervalle F3-F4 chez 5 sujets, dans l'intervalle F2-F3 chez 1 sujet, et dans l'intervalle F1-F2 chez aucun sujet).

Voyelle /a/	Tons	[+U,h]	[-U,h]	[-U,l]	[+U,l]	Voyelle /i/	Tons	[+U,h]	[-U,h]	[-U,l]	[+U,l]	Voyelle /u/	Tons	[+U,h]	[-U,h]	[-U,l]	[+U,l]
	[+U,h]		5	7	5		[+U,h]		3	4	1		[+U,h]		6	5	6
	[-U,h]			7	5		[-U,h]			1	2		[-U,h]			6	2
	[-U,l]				5		[-U,l]				2		[-U,l]				5
F3-F4	[+U,l]					F3-F4	[+U,l]					F3-F4	[+U,l]				
	[+U,h]		3	6	5		[+U,h]		3	0	2		[+U,h]		3	1	5
	[-U,h]			5	3		[-U,h]			2	4		[-U,h]			1	0
	[-U,l]				2		[-U,l]				1		[-U,l]				3
F2-F3	[+U,l]					F2-F3	[+U,l]					F2-F3	[+U,l]				
	[+U,h]		3	3	2		[+U,h]		2	0	2		[+U,h]		2	0	3
	[-U,h]			2	2		[-U,h]			0	2		[-U,h]			0	1
	[-U,l]				2		[-U,l]				1		[-U,l]				1
F1-F2	[+U,l]					F1-F2	[+U,l]					F1-F2	[+U,l]				

TABLEAU 1 : Les nombres de sujets qui font de différences significatives (sur la base de 13 sujets locutrices) en intervalles entre les formants, lorsque le ton varie (les tons concernés par la variation, marqués en rang et en colonne, sont transcrits en traits tonals) : les cas de /a/ /i/ /u/.

6 Conclusion et perspective

À travers l'analyse des données acoustiques, nous avons trouvé qu'en mandarin, le ton du segment vocalique chuchoté n'influence pas les fréquences centrales des bandes formantiques, de manière linéaire. Toutefois, ce ton a effectivement une influence sur la structure formantique, et cela se manifeste plutôt dans les intervalles entre les formants voisins qui sont, selon la littérature en psychoacoustique, associés à la sensation de la hauteur du son. Les analyses statistiques révèlent que l'intervalle entre le F3 et le F4, et celui entre le F2 et le F3 connaissent le plus souvent une modification significative ou, au moins, une tendance remarquable lorsque le ton varie. La différence maximale que le ton peut provoquer en intervalle entre les formants, calculée en *mel*, se réalise au milieu du segment, et non au début ni à la fin du segment. Ainsi, la modification de la structure formantique nous semble davantage s'orienter vers les traits tonals du type registre et non vers les traits tonals du type contour. Ce résultat corrobore, jusqu'à un certain degré, la description phonologique des tons à base de registre. Bien entendu, il est probable que le contour tonal se manifeste au moyen d'autres paramètres acoustiques, tel que l'intensité.

De plus, en comparant des cas de modification spectrale en fonction du ton de trois voyelles /a i u/, nous avons constaté que les structures formantiques de ces voyelles ont des sensibilités différentes par rapport à la variation du ton, plus précisément, dans l'ordre /a/>/u/>/i/. La modification spectrale en fonction du ton dépend également de la voyelle en question. Cela est probablement la conséquence des différences de gestes et de configurations du conduit vocal, donc au niveau supraglottique, dans la compensation de la perte de F0.

Malgré les conclusions présentées ci-dessus, l'étude est loin d'être achevée, car il nous reste deux questions principales à aborder : 1) Si l'on emploie un système tonal à base de registre, les deux couches de traits tonals [+/-U] et [h/l] auront-elles une différence ou un ordre hiérarchique dans leurs influences sur la modification de la structure formantique ? 2) Puisque la modification spectrale en fonction du ton dépend de la nature de la voyelle, quel/quels est/sont le/les aspect/aspects de cette modification qui aura/auront le poids le plus lourd dans la perception de la hauteur de la voix chuchotée ? Nous tenterons de répondre à ces questions ultérieurement avec des analyses plus avancées et des données actuelles dans la direction de la psychoacoustique, et en élaborant de nouvelles expériences qui ciblent davantage ces questions.

Références

BLICHER D. L., DIEHL R.L., COHEN L.B. (1990). Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction : Evidence of auditory enhancement. *Journal of Phonetics* 18(1), 37-49.

CHANG C., YAO Y. (2007). Tone production in whispered Mandarin. *Proceedings of the 16th ICPhS*,1085-1088. Saarbrücken, Germany.

FÓNAGY J. (1969). Accent et intonation dans la parole chuchotée. *Phonetica* 20, 177-192.

HIGASHIKAWA M., MINIFIE F. D. (1999). Acoustical-perceptual correlates of "whisper pitch" in synthetically generated vowels. *Journal of speech language and hearing research* 42, 583- 591.

KONG Y.-Y., ZENG F.-G. (2006) Temporal and spectral cues in Mandarin tone recognition. *The Journal of the Acoustical Society of America* 120, 2830-2840.

LIU S., SAMUEL A. (2004). Perception of Mandarin Lexical Tones when F0 Information is Neutralized. *Language and Speech* 47(2), 109-138.

MEYER-EPPLER W. (1957). Realization of prosodic features in whispered speech. *Acoustical Society of America* 29, 104-106.

RYALLS J. H., LIEBERMAN P. (1982). Fundamental frequency and vowel perception. *Journal of the Acoustical Society of America* 72, 1631–1634.

SEGERBÄCK B. (1965). La réalisation d'une opposition de tonèmes dans des dissyllabes chuchotés. *Studia Linguistica* 19(1-2), 1-54.

TSENG C. Y., MASSARO D. W., COHEN M. M. (1986). Lexical tone perception in Mandarin Chinese : evaluation and integration of acoustic features. *Linguistics, Psychology and the Chinese Language*. University of Hong Kong Press.

WHALEN D. H., XU Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica* 49, 25-47.

YIP M. (2001). Tonal features, tonal inventories and phonetic targets. UCL Working Papers in *Linguistics* 13, 161-188.

ZHANG J. (2014). Tones, tonal phonology and tone sandhi. *The Handbook of Chinese Linguistics*, Edited by Huang J., Li A., Simpson A. Wiley-Blackwell.

ZHANG X., SOCK R. (2015). Indices acoustiques des tons du chinois mandarin en voix normale et en voix chuchotée. *Neophilologica* 27, 266-285.

ZWICKER, E., FELDTKELLER, R., (1981) *Psychoacoustique : l'oreille, récepteur d'information*. Traduit de l'allemand par Christel Sorin, Masson.

Étude de la qualité vocale post-thyroïdectomie chez des patients souffrants ou non de paralysie récurrentielle

Ming XIU[1], Camille FAUTH[1], Béatrice VAXELAIRE[1], Jean-François RODIER[2], Pierre-Philippe Volkmar[2], Rudolph SOCK[1,3]
(1) U.R. Parole & Cognition, E.A. 1339 Linguistique, Langues et Parole (LiLPa) et Institut de Phonétique de Strasbourg (IPS) - Université de Strasbourg
(2) Groupement Hospitalier Saint Vincent – Clinique Sainte Anne – Strasbourg
(3) Language, Information and Communication Laboratory – LICOLAB, Université Pavol Jozef Šafárik, Košice, Slovaquie
ming.xiu@etu.unistra.fr

RESUME

L'objet principal de cette étude est la qualité vocale après une thyroïdectomie. Cette opération provoque souvent une dégradation de la qualité vocale de façon permanente ou temporaire. La qualité vocale sera étudiée à l'aide d'indices aérodynamiques et acoustiques.
Deux groupes de patients sont suivis et étudiés : un premier groupe de patients pour lesquels l'examen post-opératoire a révélé un défaut de mobilité de l'un des plis vocaux ; Un second groupe de patients pour lesquels l'examen post-opératoire n'a pas révélé de perturbation de la mobilité laryngée. Il s'agit d'une étude longitudinale dans laquelle la référence est constituée par la voix du locuteur en préopératoire. Les résultats préliminaires indiquent que l'ablation de la glande thyroïde modifie la voix des patients alors même que la mobilité laryngée est préservée. Tous les paramètres étudiés ont été modifiés. Le temps a toutefois un effet positif pour tous les locuteurs suivis, puisque leurs productions s'approchent, un mois après l'opération, des valeurs mesurées en préopératoire.

ABSTRACT

A post-thyroidectomy voice quality study in patients suffering or not from laryngeal paralysis
The main object of this study is voice quality after thyroidectomy. This often causes degradation of voice quality permanently or temporarily. Voice quality will be studied using aerodynamic and acoustic cues.
Data from two groups of patients were examined: one group of patients for whom postoperative examination revealed lack of mobility of one of the vocal folds; a second group of patients in whom postoperative examination revealed no perturbation of laryngeal mobility. This is a longitudinal study in which reference values are constituted by the speaker's own voice, preoperatively.
Preliminary results indicate that removal of the thyroid gland modifies the patient's voice even though laryngeal mobility is preserved. All parameters studied underwent changes. However, time has a positive effect on all speakers, since their productions resemble, one month after surgery, the values measured in the preoperative phase.

MOTS-CLES : Qualité Vocale, Thyroïdectomie, Perturbation, Paralysie/Parésie laryngée, Aérodynamique, Acoustique, Voix, Voyelles Soutenues.

KEYWORDS: Voice Quality, Thyroidectomy, Perturbation, Laryngeal Paralysis/Paresis, Aerodynamics, Acoustics, Voice, Sustained Vowels.

1 Introduction

Ce travail vise à étudier les conséquences d'une ablation totale ou partielle (isthmolobectomie) de la glande thyroïde suite à un dysfonctionnement thyroïdien, suivi ou non d'un traitement de radiothérapie. Ce type d'intervention perturbe généralement le système de production de la parole et conduit souvent à une dégradation de la qualité vocale de façon permanente ou passagère (Fauth 2012 ; Hartl et al. 2001). Cette étude entend suivre une cohorte de patients opérés de la glande thyroïde qui présente ou non une lésion de la mobilité des plis vocaux. L'étude sera longitudinale afin d'étudier les possibles stratégies de compensation ou de réajustement que le patient pourra mettre en place. Il s'agit d'étudier la flexibilité du système de production et de perception de la parole et de tenter de comprendre ce système à partir d'un dysfonctionnement d'origine pathologique (Sock & Vaxelaire. 2009).

Les deux nerfs récurrents correspondent à une des trois branches du nerf vague (nerf crânien X), qui est responsable de l'innervation des muscles laryngés intrinsèques avec le reste des nerfs innervant les muscles extrinsèques, lorsqu'il sort du tronc cérébral. Lors de l'ablation de la glande thyroïde, le chirurgien prend soin de préserver les nerfs récurrents responsables de la mobilité de tous les muscles intrinsèques du larynx, dont des plis vocaux, à l'exception du muscle crico-thyroïdien. Le nerf laryngé récurrent a un long parcours, ce qui le rend vulnérable (Aronson & Bless. 2009). La paralysie peut surprendre beaucoup de chirurgiens expérimentés de la thyroïde qui sont certains que le nerf a été laissé intact (Sulica et al. 2007). Dans de telles circonstances, il est très probable que le patient mette en place des stratégies de compensation pour contourner efficacement son trouble et tenter de conserver son intelligibilité et sa qualité vocale.

2 Protocole expérimental

Cette partie présentera le protocole expérimental que nous avons adopté pour nos travaux de recherche.

2.1 Conditions d'enregistrement et locuteurs

Le travail est effectué en collaboration avec le Groupe Hospitalier Saint Vincent et plus particulièrement avec la Clinique Sainte Anne de Strasbourg (67), où est hébergé le département de chirurgie thyroïdienne. Un soin particulier a été apporté aux conditions d'enregistrement. Il est toutefois évident que puisque les acquisitions de données ont eu lieu en milieu hospitalier, elles ne peuvent pas être optimales. Les patients ont été enregistrés durant toutes les phases dans un endroit calme, lors des consultations ce qui explique le choix d'un corpus restreint. Le premier enregistrement a eu lieu le jour précédant l'opération, et les suivants ont eu lieu après l'opération, à J + 1 jour, J + 15~20 jours et J + 1 mois.

Il y a deux groupes de locuteurs. Le premier groupe de locuteurs est composé de neuf locuteurs ayant subi une opération de la glande thyroïde et ne présentant pas d'immobilité laryngée. Le deuxième groupe de locuteurs est composé de 3 locuteurs présentant une immobilité laryngée post-thyroïdectomie ; un locuteur présente une paralysie laryngée de la corde vocale droite en adduction et deux locuteurs présentent une parésie laryngée également de la corde vocale droite,

mais en position d'abduction. Tous nos locuteurs sont de langue maternelle française et ne présentaient aucune pathologie vocale ou auditive avant l'opération chirurgicale.

La différence entre la paralysie et la parésie est que la parésie correspond à une perte partielle des capacités motrices d'une ou deux cordes vocales de façon transitoire, et que la paralysie provoque la perte totale de motricité d'une ou deux cordes vocales de façon transitoire ou permanente (Venkatesan. 2011).

2.3 Corpus et méthode

Afin d'obtenir des données aérodynamiques et acoustiques, deux méthodes d'acquisition différentes ont été utilisées.

Il s'agissait d'inspirer profondément avant de tenir le plus longtemps possible, à une hauteur et une intensité confortable, la voyelle /a/ afin de mesurer le temps maximum phonatoire. Cette tâche a été répétée deux fois afin de permettre une acquisition en aérodynamique, puis en acoustique.

Afin d'obtenir des données aérodynamiques, nous avons utilisé la plateforme EVA2 (Ghio & Teston 2004). Le débit d'air oral (Oaf) a ainsi pu être obtenu ; cette mesure illustre les stratégies au niveau aérodynamique pendant la phonation. Nous avons également pu mesurer ainsi le temps maximum phonatoire.

Les données acoustiques ont été enregistrées grâce à un enregistreur numérique Marantz Professional© PMD661, avec microphone Sennhaiser© e 835. Lors de l'enregistrement, le sujet était assis sur un tabouret haut à environ 20 cm du microphone. Les enregistrements acoustiques ont permis de mesurer la fréquence fondamentale, le jitter et Harmonics to Noise Ratio.

3 Hypothèses

Nous formulons les hypothèses suivantes :

1. La voix du patient, sans atteinte de la mobilité laryngée (groupe SP), pourrait se trouver modifier dans les phases post-opératoires précoces. Au niveau acoustique, la modification de la voix pourrait affecter les valeurs de la fréquence fondamentale (F0). L'activité irrégulière au niveau du larynx pourrait avoir des conséquences sur les mesures de perturbations du signal, augmentant le jitter et abaissant les valeurs de Harmonics to Noise Ratio (HNR).
2. Au niveau aérodynamique, les patients sans atteinte de la mobilité laryngée (groupe SP), pourraient avoir des difficultés à atteindre les valeurs de temps maximum phonatoire (TMP) enregistrées en préopératoire. De plus, le débit d'air oral (ou Oaf) pourrait être un indice permettant d'illustrer les efforts respiratoires que les patients doivent fournir en post-opératoire pour conserver l'efficacité vocale.
3. La section du nerf récurrent est responsable de déficits moteurs au niveau de l'innervation des muscles laryngés. En conséquence, chez les patients présentant une paralysie laryngée (Groupe AP), toutes les modifications précédemment évoquées seront plus importantes et probablement plus persistantes.

4. Enfin, le temps devrait permettre aux patients ne présentant pas de paralysie laryngée de retrouver rapidement les valeurs obtenues, lors de l'enregistrement préopératoire. En revanche, pour

les patients présentant une immobilité laryngée, cette récupération sera plus lente et devrait être accompagnée d'une rééducation orthophonique.

4 Résultats

Nous présentons ci-après les résultats obtenus à partir des deux groupes de locuteurs. Dans le cas du groupe AP, nous avons choisi de présenter les valeurs de façon individuelle puisque les locuteurs de ce groupe n'avaient pas le même type d'atteinte de la mobilité laryngée. Pour les locuteurs du groupe SP, lorsque cela était possible, nous présentons les valeurs moyennes. Les valeurs entre parenthèse renvoient aux écarts-types.

4.1 Données aérodynamiques

Après une thyroïdectomie, une surconsommation d'air pourrait être observée directement à partir des données aérodynamiques sur les valeurs d'Oaf, ce qui pourrait être une explication de la diminution du TMP.

L'Oaf a été modifié pour tous les sujets du groupe SP (voir figure 1) après l'opération. La valeur moyenne a diminué de 0.151 dm^3/s (0.04) en Préop jusqu'à 0.135 dm^3/s (0.06) en PO1. Dans la deuxième phase post-opératoire (PO2), la plupart des sujets ont montré une augmentation de la valeur de ce paramètre, soit 0.162 dm^3/s (0.06) en moyenne. Toutefois, L'Oaf est redevenu comparable aux mesures préopératoires en PO3, soit 0.156 dm^3/s (0.06) en moyenne.

Sur la figure 2, sont représentées les valeurs pour le groupe AP. Pour le sujet 1 qui présente une parésie, on observe d'abord un abaissement en PO1 (0.112 dm^3/s par rapport à la phase préopératoire : 0.126 dm^3/s). Puis, une augmentation importante apparait dans la deuxième phase post-opératoire (PO2), où la valeur d'Oaf est de 0.374 dm^3/s. Ce paramètre d'Oaf est de 0.229 dm^3/s en PO3 ; il tend alors à s'approcher des valeurs mesurées en préopératoire. Pour le sujet 4 qui présente une paralysie, nous avons observé une augmentation importante de son Oaf dans la première phase post-opératoire (PO1), mesuré à 0.664 dm^3/s (0.291 dm^3/s en préopératoire). Puis, cette valeur baisse progressivement dans les deux phases suivantes, 0.496 dm^3/s en PO2, 0.284 dm^3/s en PO3. Il s'approche alors du niveau mesuré en préopératoire, soit 0.291 dm^3/s. En ce qui concerne le sujet 10, qui présente aussi une parésie (figure 2), un abaissement est apparu en PO1 (0.101 dm^3/s par rapport à la phase préopératoire : 0.135 dm^3/s). S'en suit une augmentation légère qui apparait dans la deuxième phase post-opératoire (PO2), où la valeur d'Oaf est de 0.125 dm^3/s, pour s'approcher des valeurs mesurées en préopératoire. Son paramètre d'Oaf est finalement revenu à 0.127 dm^3/s en PO3.

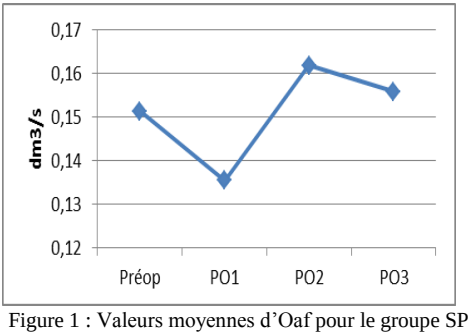


Figure 1 : Valeurs moyennes d'Oaf pour le groupe SP

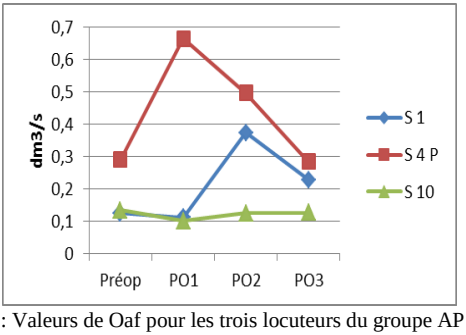


Figure 2 : Valeurs de Oaf pour les trois locuteurs du groupe AP

Pour la plupart des patients du groupe SP (figure 3), on observe une augmentation du temps maximum phonatoire en post-opératoire 1, où il est mesuré à 12.07 sec (4.12). Cette valeur a continué à augmenter progressivement en post-opératoire 2 (12.5 sec (3.79)) et post-opératoire 3 (13.87 sec (4.41)), où les valeurs obtenues dépassent alors celles mesurées en phase préopératoire (11.86 sec (2.99)). Notons que pour toutes les phases d'enregistrement, les écarts-types restent importants, ce qui révèle une certaine variabilité dans ces données pathologiques.

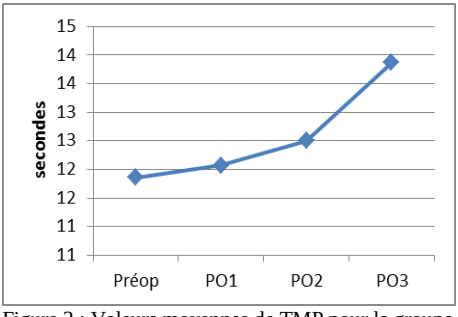


Figure 3 : Valeurs moyennes de TMP pour le groupe SP

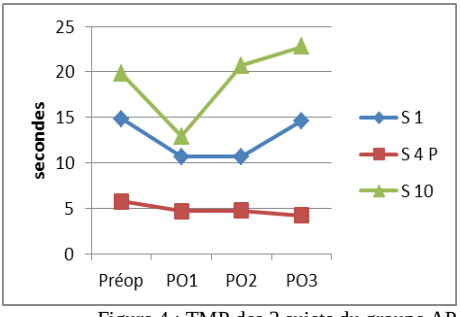


Figure 4 : TMP des 3 sujets du groupe AP

Pour les sujets du groupe AP (figure 4), on constate un pattern différent. En premier lieu, notons que le sujet 4 avait avant l'opération déjà un temps maximum phonatoire faible (5.7 sec). Après l'opération, il passe de 4.7 sec (en PO1 et PO2) à 4.2 sec en post-opératoire 3. Le temps n'a donc pas eu d'effet positif sur ce paramètre, chez ce locuteur. Chez les sujets 1 et 10, le temps maximum phonatoire est diminué après l'opération, par rapport aux valeurs obtenues en préopératoire. Le temps a un effet positif sur ce paramètre, puisque les valeurs augmentent dès la phase post-opératoire 2 pour le sujet 10, et à partir de la phase post-opératoire 3, pour le sujet 1. Lors de la dernière phase, les valeurs sont alors comparables à celles mesurées en préopératoire.

4 .2 Données acoustiques

Il est judicieux de penser que la modification aérodynamique pourrait affecter les indices acoustiques. Cette modification de la voix influencerait directement les valeurs de F0. Dans la mesure où ce paramètre est directement lié à l'activité laryngienne, nous avons trouvé un abaissement de la F0 dans les phases postopératoires précoces pour la plupart des patients comme illustré sur la figure 6.

Pour les locuteurs du groupe AP, les résultats sont à considérer en fonction des locuteurs. Pour le sujet 4 (sujet masculin), la fréquence fondamentale est mesurée à 98,7 Hz en préopératoire, puis à 124 Hz, 140,3 Hz et 101,5 Hz en post-opératoire 1, 2 et 3 respectivement. Le sujet 1 a montré aussi une augmentation en PO1 comme le sujet 4. Il a eu d'abord une augmentation en PO1, (232.8 Hz par rapport à la phase préopératoire : 163.7 Hz). Puis, un abaissement apparait en PO2 où la valeur de F0 est de 191.2 Hz. En PO3, sa valeur est comparable à celle mesurée en PO2, soit à 190.88 Hz. Pour le sujet 10, la fréquence fondamentale baisse en post-opératoire 1 (217.7 Hz), par rapport à celle mesurée en préopératoire (236.6 Hz). À partir de la phase post-opératoire 2, cette mesure augmente à 298.8Hz et à 345.1 Hz en post-opératoire 3.

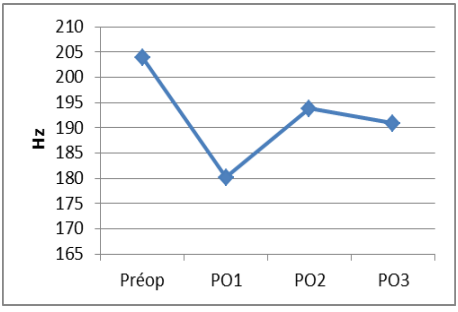


Figure 6 : Valeurs moyennes de F0 pour le groupe SP

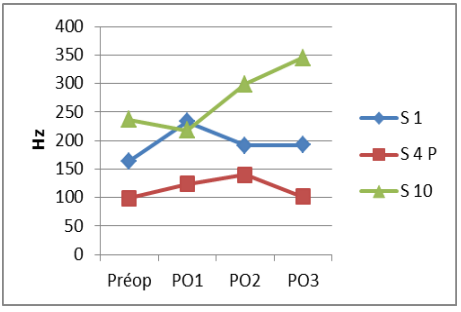


Figure 7 : F0 des 3 sujets du groupe AP

Nous pouvons observer que l'indice du jitter (figure 8) a un abaissement léger dans la phase PO1 pour les patients du groupe SP (figure 8 et 10).

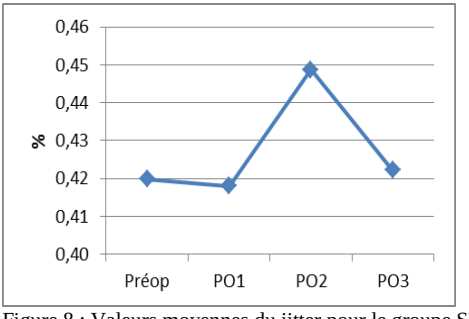


Figure 8 : Valeurs moyennes du jitter pour le groupe SP

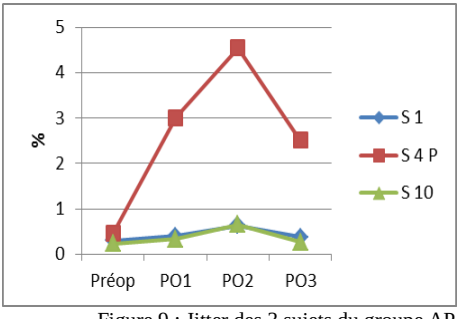


Figure 9 : Jitter des 3 sujets du groupe AP

En ce qui concerne les patients ne présentant pas d'immobilité laryngée, le jitter moyen a abaissé de 0.42% (0.084) en Préop à 0.418% (0.153) en PO1. Contrairement aux paramètres précédemment étudiés, le jitter ne se normalise pas dès la phase post-opératoire 2 ; il continue à augmenter sans pour autant être très éloignés des valeurs standard attendues (0,449% (0.16) pour le jitter). En phase post-opératoire 3, ces indices sont comparables à ceux obtenus en phase préopératoire (0.422% (0.084) pour le jitter). Le sujet 4 a des valeurs plus remarquables pour le jitter. La valeur de son jitter en phase Préopératoire était de 0.46%. On observe ensuite chez lui une augmentation très importante à 3% en PO1. Puis, il a continué à augmenter son jitter à 4.5% en PO2. Dans la dernière phase (PO3), il le baisse à 2.5%. Les sujets 1 et 10 restent tout au long des phases d'enregistrement proches des valeurs normales pour ce qui concerne ces deux paramètres.

Nous pouvons observer une diminution du HNR (figure 10) pour le groupe SP. Avant l'opération, le HNR moyen était de 19.27 dB (1.57), il est ensuite de 18.91 dB (1.81) en PO1, de 17,9 dB (2.69) en PO2 et de 17.62 dB (1.82) en PO3. Notons que les écarts types sont faibles.
C'est naturellement le sujet 4 (figure 11), qui présente une immobilité laryngée, pour qui le paramètre du HNR est significativement modifié après l'opération. En préopératoire, la mesure du HNR est proche des valeurs standard à 17,91 dB. Après l'opération, ce paramètre chute à 1.61 dB en PO 1 et 1.36 dB en PO2. L'augmentation de la valeur est légèrement amorcée en post-opératoire 3, en remontant à 6.27 dB.

C'est encore, comme attendu, le sujet 4 (figure 11) qui présente une immobilité laryngée, pour qui le paramètre du HNR est significativement modifié après l'opération. En préopératoire, la mesure du HNR est proche des valeurs standard à 17,91 dB. Après l'opération, ce paramètre chute à 1.61

dB en PO 1 et 1.36 dB en PO2. L'augmentation de la valeur est légèrement amorcée en post-opératoire 3, remontant à 6.27 dB.

En ce qui concerne les sujets 1 et 10 (figure 11), on observe une diminution de leur HNR en post-opératoire 1 (18.17 dB et 22.4 dB respectivement), par rapport aux mesures obtenues en préopératoire (20.2 dB et 23.02 dB, respectivement). Cette diminution se confirme en post-opératoire 2 (14.24 dB et 17.41 dB respectivement). Ce paramètre augmente à partir de la phase postopératoire 3 (18.3 dB et 20.7 dB respectivement) pour s'approcher des valeurs mesurées en préopératoire.

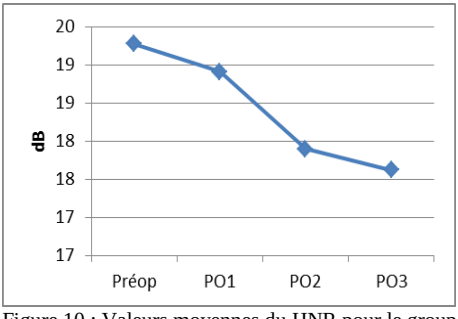


Figure 10 : Valeurs moyennes du HNR pour le groupe SP

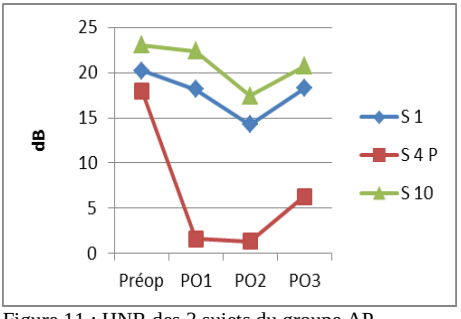


Figure 11 : HNR des 3 sujets du groupe AP

5 Discussion

Aussi bien dans la dimension aérodynamique que dans la dimension acoustique, les résultats confirment partiellement nos hypothèses.

En ce qui concerne les mesures acoustiques (H1), nous observons effectivement une modification des mesures de perturbations du signal (jitter et HNR), dans les phases d'enregistrement post-opératoires 1 et 2. Dès la phase d'enregistrement post-opératoire 3, les valeurs pour les patients SP sont comparables à celles que nous avions mesurées avant l'opération. Par ailleurs, l'abaissement du HNR peut être analysé en regard de l'augmentation du jitter. Pour les patients AP, la normalisation de ces paramètres est amorcée. La F0 a été modifiée après l'opération pour tous les sujets. Dans la dernière phase postopératoire, même si les valeurs des paramètres évoluent favorablement, ils restent tout de même relativement déviants par rapport aux valeurs préopératoires, sauf pour un sujet qui présente une parésie.

En ce qui concerne les mesures aérodynamiques (H2), nous avons observé une modification de l'Oaf. Il est possible que les patients mettent en place une stratégie de compensation qui vise à pallier les activités glottiques irrégulières en « surconsommant » de l'air. L'instabilité glottique et la mauvaise gestion pneumo-phonatoire sont probablement responsables de la diminution du TMP. L'immobilité laryngée accentue cette observation. Le sujet est obligé de surconsommer de l'air à partir du même volume pulmonaire. Avec le temps, les patients ont retrouvé des stratégies efficaces pour l'exploitation de la colonne d'air subglottique.

Enfin, il a été possible de montrer une corrélation entre les données aérodynamiques et les données acoustiques. Pour les patients qui présentent une paralysie laryngée (H3), les modifications sont plus importantes et aussi plus persistantes, à la fois au niveau aérodynamique qu'acoustique. Que l'on considère les patients avec ou sans paralysie laryngée, le temps a un effet bénéfique sur les tous les

paramètres étudiés, qui évoluent vers les valeurs mesurées avant l'opération. La récupération est simplement plus lente (H4) pour les patients qui présentent un défaut de mobilité de l'un des plis vocaux. Notons que les modifications pour les patients sans paralysie sont pour la plupart non significatives. Ces résultats peuvent être rapprochés de ceux observés dans la littérature.

6 Conclusions

De façon générale, nous observons un schéma identique pour tous les paramètres étudiés. En ce qui concerne le groupe sans immobilité laryngée, les paramètres sont modifiés dans les phases post-opératoires 1 et 2. En ce qui concerne les mesures aérodynamiques, les valeurs sont plus perturbées en post-opératoire 2 qu'en post-opératoire 3. Il pourrait s'agir d'une phase de réadaptation, en effet après l'opération la perturbation de la qualité vocale peut être provoquée par de multiples causes (intubation prolongée, position du cou en hyper extension pendant l'opération, ablation d'un goitre…). Les écarts-types sont plus importants dans les phases post-opératoires, témoignant de la forte variabilité intra-locuteurs et interlocuteurs dans ces phases. Les patients présentant un défaut de mobilité laryngé présentent un schéma identique, les modifications sont en revanche plus importantes. En ce qui concerne les patients présentant un défaut de mobilité de l'un des plis vocaux, il est clair que la position de la corde vocale (en adduction ou en abduction) et la gravité de l'atteinte nerveuse jouent un rôle prépondérant sur la dégradation de la qualité vocale.

Toutes ces tendances seront à confirmer par l'utilisation de tests statistiques afin de vérifier la robustesse de nos analyses. Nous cherchons actuellement à établir diverses corrélations potentielles entre les différents paramètres utilisés.

Remerciements

Ce travail a été financé par un projet IdEx de l'Université de Strasbourg 2014 – Attractivité «Plateforme Unistra de Phonétique et Linguistique Cliniques» porté par B. Vaxelaire. Les auteurs remercient en outre tous les locuteurs qui ont accepté de participer à cette étude.

Références

ARONSON A., & BLESS D. (2009). *Clinical Voice Disorders, Book + DVD* (p. 366-367). Thieme.

BAUJAT B, DELBOVE H, WAGNER I., FUGAIN C., DE CORBIERE S. & CHABOLLE F. (2001). Immobilité laryngée post thyroïdectomie. *Annales de Chirurgie*, 126(2), 104-10.

DEBRUYNE F., OSTYN F., DELAERE P., & WELLENS W. (1997). Acoustic analysis of the speaking voice after thyroidestomy. *Jounal of voice: Official journal of the Voice Foundation*, 11(4), 479-482.

FAUTH C. (2012). *Perturbation de la production de la parole suite à une opération de la glande thyroïde*. Thèse dedoctorat en Phonétique Générale et Expérimentale, Phonétique Clinique. Strasbourg : Université de Strasbourg. 2012. 316p.

GHIO A., & TESTON B. (2004). Evaluation of the acoustic and aerodynamic constraints of a pneumotachograph for speech and voice studies. *International Conference on Voice Physiology and Biomecanics*, Marseille, France, August 18-20, 2004, p.55-58

HARTL D., HANS S., VAISSIERE J., RIQUET M., & BRASNU D. (2001). Objective Voice Quality Analysis Before and After Onset of Unilateral Voval Fold Paralysis. *Jounal of Voice, 15(3)*, 351-361.

JACOBSON B, JOHNSON A., GRYWALSKI C. SILBERGLEIT A., JACOSBON G, BENNINGER M.S., NEWMAN C.W. (1997). "The voice handicap index (VHI): development and validation," *J.Speech-Lang.Path.*, vol. 6, pp. 66-70.

KEILMANN A., & HÛLSE M. (1992). Dysphonie nach Strumektomie bei ungestörter respiratorischer Beweglichkeit der Stimmlippen. *Folia Phoniatrica et Logopaedica*, *44*(6), 261-268.

SOCK R., & VAXELAIRE B. (2009) How special is speech? *In* "Some Aspects of Speech and the Brain". S . FUCHS H. LOEVENBRUCK D. PAPE P. PERRIER (Eds.). Peter Lang Internationaler Verlag der Wissenschaften, Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien, 2009, 259-294.

SULICA L., & BLITZER A. (2007) *Vocal fold paralysis* (p. 38-42). Springer.

VENKATESAN N. (2011). *Unilateral vocal fold paralysis*. Cours du Département d'Oto-rhino-laryngologie. The University of Texas Medical Branch. Texas.

Etude par EMA des mouvements de la mâchoire inférieure durant les consonnes de l'arabe marocain

Chakir zeroual[1, 2] Phil Hoole[3] Adamantios Gafos[4]
(1) Faculté Polydisciplinaire de Taza, BP. 1223, Taza, Maroc.
(2) Laboratoire de Phonétique et Phonologie, CNRS-UMR7018, Paris, France
(3) Institut fuer Phonetik und Sprachverarbeitung, University of Munchen, Germany
(4) University of Potsdam, Germany

chakirzeroual@yahoo.fr, hoole@phonetik.uni-muenchen.de,
adamantios.gafos@uni-potsdam.de

RESUME

Cette étude est basée sur des données obtenues à l'aide d'EMA (AG500) enregistrant les mouvements de la mâchoire inférieure (Minf) durant les consonnes labiales, coronales, vélaires, uvulaires, pharyngales et laryngales de l'arabe marocain dans les contextes aCa et iCi. Nous avons montré que l'implication de la Minf est cruciale durant /s S t T/ (S T : consonnes emphatiques). Le recul de la racine de la langue n'est pas nécessairement corrélé à la baisse de la Minf. Les consonnes apicales ne sont pas toujours associées à l'abaissement de la Minf. La Minf ne semble pas impliquée durant les laryngales et les pharyngales, ce qui est en accord avec les déductions de Goldstein (1995). Les mouvements verticaux et horizontaux de la Minf sont relativement indépendants.

ABSTRACT

EMA study of jaw movements during Moroccan Arabic consonants
Our observations suggest that the degree of jaw involvement is crucial during coronal voiceless obstruent /s S t T/ (S T : pharyngealized cosonants). /t T/ have similar height even though /t/ is laminal and /T/ apical: Apicality seems not always correlated with jaw lowering. The high jaw position during /T/ seems necessary to support high tongue tip position when the tongue root is strongly retracted. The jaw seems not activated during the pharyngeal and laryngeal consonants: in accord more with Goldstein (1994) than Elgendy (1999) hypothesis. High/low and front/back jaw movements are not always correlated as is expected from a simple rotation movement of the jaw.

MOTS-CLES : EMA, mâchoire inférieure, production de la parole, arabe marocain,.
KEYWORDS: EMA, jaw, speech production, Moroccan Arabic.

1 Introduction

Peu d'études articulatoires ont essayé de caractériser les propriétés spatiotemporelles de la mâchoire inférieure (Minf) durant les consonnes. Elles ont généralement porté sur les labiales, coronales et/ou vélaires, et ont montré que les propriétés articulatoires de la Minf varient en fonction de leur lieu et mode d'articulation et sont associées, en partie, à des contraintes biomécaniques combinées éventuellement à des contraintes perceptives. Ici, nous nous baserons sur les études de Keating et al.

(1994 : anglais et suédois), Lee et al. (1995 : français, coréen, arabe), Mooshammer et al. (2006 et 2007 : allemand), Recasens et al. (2012 : espagnol), Elgendy (1999 : arabe égyptien), Lindblom (1983 : suédois) et Zeroual et al. (2007 : arabe marocain, désormais AM).

/s ʃ/ ont la position la plus élevée de la Minf qui reste quasi-stable quelles que soient les voyelles adjacentes (Keating et al., 1994 ; Lee et al., 1995 ; Mooshammer et al. 2007). Cette posture est attribuée à la précision articulatoire requise durant ces consonnes, ainsi qu'à la présence d'une source supplémentaire du bruit de friction entre les incisives inférieures et supérieures (Shadle, 1985) qui renforce leurs caractéristiques acoustiques. La Minf est plus avancée durant /ʃ/ comparée à /s/ (Mooshammer et al. 2007) liée très probablement à l'arrondissement durant la première.

/t/ (avec VOT long) a également une position élevée et invariable de la Minf ; sa hauteur est similaire (Keating et al. 1994) ou plus basse comparée à /s/ (Lindblom, 1983 ; Mooshammer et al. 2007). Selon Mooshammer et al. (2006), la montée substantielle de la Minf durant /t/ est nécessaire pour produire un bruit de relâchement long et saillant. La montée maximale de la Minf est souvent alignée avec le relâchement de /t/, mais située bien avant durant /d/ (Mooshammer et al., 2006 ; Zeroual et al. 2007). Mooshammer et al. (2007) montrent que la Minf est plus basse durant /d/ comparée à /t/ qu'ils associent au fait que /d/ peut être apicale et /t/ laminale (voir aussi Dart, 1991).

Comparées aux autres coronales notamment obstruantes, /l r/ (apicales) ont souvent une position plus basse de la Minf qui semble nécessaire pour éviter, durant /l/, un contact latéral entre la langue et le palais (Mooshameer et al., 2006) et pour faciliter, durant /r/, la montée et la rétraction de la pointe de la langue. La hauteur de la Minf durant /l/ peut être supérieure (Elgendy, 1999), identique (Keating, 1994), ou plus basse (Lindblom, 1983) comparées à /k g/.

/p b/ ont généralement une hauteur la Minf entre les dentales et les vélaires : /t, d/ > /p, b/ > /k, g/ (Keating et al., 1994; Lee, 1995). La position abaissée de Minf durant /p b/, comparée à /t d/, est due aux lèvres qui peuvent se déplacer relativement indépendamment de la Minf. Keating et al. (1994) et Elgendy (1999) ont rapporté une position plus élevée de la Minf durant /f/ comparée à /b/ liée très probablement à la constriction labiodentale. /k g/ sont généralement associées à une position plus basse de la Minf, comparées aux obstruantes labiales et coronales, qui coarticule assez fortement avec les voyelles adjacentes. La position abaissée de la Minf, durant /k g/, est généralement attribuée à son axe de rotation qui est plus proche de l'articulateur dorsal (Hoole and Kühnert, 1996, Keating et al. 1994). En effet, le contact dorso-vélaire ne nécessite pas une montée importante de la Minf et la rotation de cette dernière affecte faiblement la hauteur du dos de la langue.

Très peu d'études ont été consacrées aux consonnes post-vélaires, celle de Boff (1983, ciné-radiographie) a montré que, dans aCa, les pharyngales de l'AM affichent la baisse la plus importante de la Minf, qui est plus marquée que /a/ (réalisée [æ] en AM). Viennent ensuite les laryngales, puis les uvulaires. Une gradation similaire a été rapportée par Elgendy (1999) pour qui l'abaissement de la Minf durant les pharyngales serait actif et permettrait à la racine de la langue de reculer plus facilement. Pour Nolan (1995), cet abaissement de la Minf est une conséquence passive de la montée du larynx durant les pharyngales. La coarticulation très marquée de la Minf avec les voyelles, durant les laryngales et les pharyngales, combinée au fait que les pharyngales ont en arabe iraquien une position plus élevée de la Minf et plus abaissée en AM, ont amené Goldstein (1994) à suggérer que les laryngales et les pharyngales sont produites sans implication active de cet articulateur. En fait pour Goldstein (1994), toutes les gutturales /χ ʁ ħ ʕ h ʔ/, qui constituent une classe naturelle, serait caractérisée par la non implication active de la Minf.

Cette étude exhaustive, basée sur les données de 3 locuteurs produisant plusieurs consonnes de l'AM (Table 1), teste une partie des hypothèses articulatoires et perceptives citées ci-dessus : (i) le bruit de friction saillant des coronales suppose nécessairement une position très élevée de la Minf ; (ii) les consonnes apicales ont une position plus basse de la Minf comparées à leurs correspondantes laminales ; (iii) la rétraction de la racine de la langue nécessite un abaissement de la Minf ; (iv) les gutturales sont produites sans implication active de la mâchoire inférieure. Notons que plusieurs critères peuvent montrer l'implication active de la Minf (Goldsetin, 1994, Moosammer et al, 2006 et 2007). Nous nous baserons ici sur celui qui consiste à comparer, dans les mêmes contextes vocaliques, les positions spatiales de la Minf durant la production des consonnes ayant des lieux et des modes d'articulation différents. Nous analyserons également l'ampleur de l'effet (coarticulation) des voyelles sur les mouvements de la Minf durant ces consonnes.

2 Méthodologie

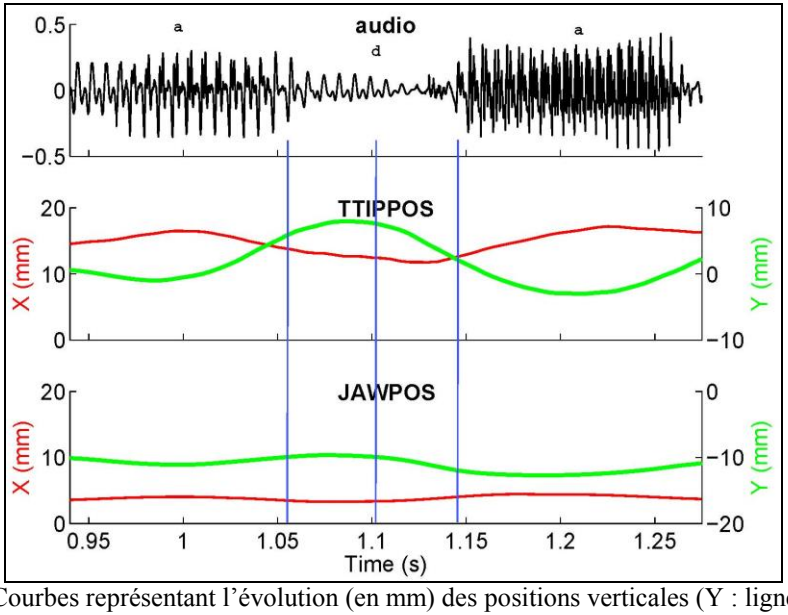


Figure. 1 : Courbes représentant l'évolution (en mm) des positions verticales (Y : lignes vertes) et horizontales (X : lignes rouges) de la pointe de la langue (TTIPPOS) et de la mâchoire inférieure (JAWPOS) durant /ada/ dans /madabʃ/. Les lignes verticales bleues correspondent aux postions (de gauche à droite) onset, médiane et offet de /d/ (/C/), où les mesures (x et y) ont été relevées.

4 locuteurs natifs de l'AM ont participé à une expérience par EMA 3-dimensionnelle (AG500, Carstens Medizinelektronik). Cette technique nous a permis d'enregistrer les mouvements de plusieurs articulateurs grâce à des capteurs fixés : proche de la pointe de langue, au niveau de sa partie médiane, sur son dos, sur les extrémités externes centrales des lèvres inférieures et supérieures et sur la base externe des incisives inferieures pour enregistrer les mouvements de la Minf.

Un programme Mview, développé sur Matlab par M. Tiede (Haskins), nous a permis d'identifier automatiquement, à partir de la courbe de la vélocité (seuil de 20%), les positions temporelles et spatiales des mouvements de chaque articulateur, ainsi que les valeurs de la vélocité maximale et l'amplitude de ses phases de fermeture et d'ouverture. Grâce à ce programme, nous avons également

relevé les mesures de la position verticale et horizontale de la Minf au niveau de l'onset, partie médiane, et l'offset des trois segments des suites iCi et aCa. Cette étude sera basée essentiellement sur l'analyse des mesures effectuées au niveau de la partie médiane de /C/ (Figure 1).

/C/		/ma+CaC+ʃ/	/y+CiC/	Sens
/b/	Labiale occlusive	[mabanʃ]	[ibiʕ]	Apparaître ; vendre
/d/	Coronale occlusive non emphatique	[madabʃ]	[idir]	Fendre ; faire
/D/	Corononale occlusive emphatique	[maDamʃ]	[iDim]	Damer (un pion)"
/t/	Coronale occlusive non emphatique	[matabʃ]	[itih]	Se repentir
/T/	Coronale occlusive emphatique	[mɑTɑbʃ]	[iTib]	Cuire
/s/	Coronale fricative non-emphatique	[masabʃ]	[isib]	Abandonner
/S/	Coronale fricative emphatique	[mɑSɑbʃ]	[iSib]	Trouver
/l/	Coronale latérale	[malamʃ]	*[iliz]	Prendre le parti de quelqu'un
/r/	Coronale battue (ou tap consonant)	[marabʃ]	[iɾib]	s'effondrer
/k/	Vélaire occlusive	[makanʃ]	*[ikis]	Etre
/g/	Vélaire occlusive	[magalʃ]	[igis]	Dire ; Essayer
/q/	Uvulaire occlusive	[maqasʃ]	[iqis]	Toucher
/χ/	Uvulaire fricative	[maχafʃ]	[iχib]	Avoir peur ; Etre déçu
/ħ/	Pharyngale approximante	[maħalʃ]	[iħil]	Ternir
/h/	Laryngale fricative	[mahamʃ]	[ihim]	S'égarer

TABLE 1 : Consonnes de l'AM concernées par cette étude et placées dans les contextes aCa et iCi. Emphatique renvoie à une articulation secondaire caractérisée comme une pharyngalisation ou uvularisation, d'où les symboles non standards /T D S/. /t/ a un VOT très long et une articulation laminale ; /T/ un VOT très court et une articulation apicale (Zeroual et. 2007). /r/ est réalisé [ɾ]. /a/ est phonétiquement [æ] sauf au voisinage d'une emphatique où elle est réalisée [ɑ]. * : non mot.

Des items de l'AM (Table 1) complétés par quelques non mots et classés de manière aléatoire ont été prononcés (8 fois) par 4 locuteurs natifs dans la phrase cadre : [galha ____ hnaja] « il lui a dit __ ici ». Dans ces items, plusieurs types de consonnes apparaissent dans les contextes symétriques aCa et iCi. Ces items sont des verbes conjuguées à la forme accomplie /ma+CaC+ʃ/ et inaccomplie /yCiC/, 3ème personne masculin singulier. /yCiC/ est réalisé phonétiquement [iCiC]. Dans /ma+CaC+ʃ/, [ma…ʃ] est un morphème discontinu de négation. Dans tous ces items, l'accent est placé la 2ème voyelle. Cette étude est limitée aux données de 3 locuteurs réalisant 5 répétées par item.

3 Résultats et discussions

Une ANOVA à deux facteurs montrent que la hauteur de la Minf varie significativement en fonction de la nature de la consonne [$F(14 ; 420)=37,1$; $p<0,001$] et du contexte vocalique [$F(1 ; 420)=62,7$; $p<0,001$]) ; leur interaction est significative ($p<0,001$). Une deuxième ANOVA à 2 facteurs montre que la position horizontale de la Minf varie en fonction de la consonne [$F(14 ; 420)=25,35$; $p<0,001$] et du contexte vocalique [$F(14 ; 420)=17,94$; $p<0,001$] ; leur interaction est non significative ($p=0,23$). Les valeurs moyennes des positions verticales et horizontales de la Minf sont données dans la Table 3, les analyses post-hoc (TukeyHSD) de la première ANOVA sont résumés dans la Table 4 (positions verticales) et de la deuxième dans la Table 5 (positions horizontales).

3.1 Mouvements verticaux de la mâchoire inférieure en fonction de /C/

Nos résultats montrent que /t d T D s S/ ont la même position verticale de la Minf qui est significativement supérieure aux autres consonnes /b l r k g q χ ħ h/ (seule /d/ vs /b/ est non significative). Ces résultats semblent confirmer l'hypothèse selon laquelle la production d'un bruit de friction très saillant durant /s S/ et /t/ (qui a VOT long) nécessite une montée substantielle de la Minf. La hauteur identique de la Minf durant les emphatiques /T D S/, comparées à leurs correspondantes non emphatiques /t d s/, montre que le recul de la racine de la langue, pour produire leur articulation secondaire dans la cavité pharyngale, n'est pas nécessairement corrélé à la baisse de la Minf. Nous pensons que la montée importante de la Minf durant /T D S/, mêmes si elles sont apicales, est pour compenser le recul de la racine de la langue due à leur pharyngalisation.

Le fait que les coronales obstruantes ont une position beaucoup plus élevée de la Minf comparées à /l r/ est en accord avec les observations de plusieurs études. /l r/ ont la Minf légèrement plus basse mais non statiquement différente de /k g/. Notons que seule Lindblom (1983 : suédois) rapporte une position significativement plus basse de la Minf durant /l r/ comparées à /k g/. Les observations de Keating (1994) montrent une position identique de la Minf durant /l r k g/ avec la tendance [l r] < [k g] (pour l'anglais). Dans les données d'Elgendy (1999), la Minf durant /l r/ est significativement plus élevée comparée à /k g/. Dans nos données, la position légèrement plus haute de la mâchoire inférieure durant /k g/ comparée à /l r/ peut être due au fait que /k/ de l'AM est dorso-palatal devant /a, i/ (Boff, 1983) et que /a/ de l'AM est réalisé [æ] sauf au voisinage d'une emphatique. Zeroual et al. (2011a et 2011b) montrent aussi que /k/ de l'AM est produit devant /a i/ alors que le dos de la langue est dans la même position horizontale qui est plus avancée comparée à sa position devant /u/.

/b/ est produite avec une position de la Minf qui est plus élevée, mais non significativement, à celle de /l r k g/. Cette tendance s'accorde avec les travaux qui montrent que la hauteur de la Minf durant /p b/ suie la gradation suivante : /t d/ > /p b/ > /k g/ (Keating et al., 1994; Lee, 1994). La Minf est plus élevée durant /k g/ comparée à /q χ ħ h/ ; seules les différences /k g/ vs. /q/ sont non significatives. Les observations par EMA de Zeroual et al. (2011a) montrent que /q/ est très palatalisée dans iCi, ce qui peut expliquer la hauteur très proche de la Minf durant /k g q/.

Les gutturales /χ ħ h/ ont la position la plus abaissée de la Minf comparées à toutes les autres consonnes. Seules /ħ h/ présentent la position de la Minf qui est statiquement abaissée comparée à toutes les autres consonnes buccales. Même si la Minf est plus basse durant /q χ/ comparée à /l r/, cette différence n'est pas statistiquement significative. Nous pensons que ce comportement particulier de /q χ/, se rapprochant à la fois des consonnes buccales et des gutturales, est lié au fait qu'ils sont des « segments complexes » ayant une articulation buccale impliquant le dos de la langue et une autre pharyngale réalisée par la racine de la langue (Goldstein, 1994).

3.2 Mouvements horizontaux de la mâchoire inférieure en fonction de /C/

/s S/ ont une position de la Minf qui est la plus avancée comparée à toutes les autres consonnes, y compris /d T D/ (seule /t/ vs. /s S/ non significative). /t/ est également plus avancée que le reste des consonnes à l'exception de /d/. Rappelons que /s S t d T D/ sont produites avec une hauteur de la Minf qui est statistiquement similaire. Ces résultats suggèrent que le mouvement de translation horizontale de la Minf peut être contrôlé indépendamment de son mouvement de rotation.

/s/ et /d/ ont une même position horizontale comparée respectivement à leur correspondante emphatique /S/ et /D/. Ces deux résultats, montrent que le recul de la racine de la langue durant les consonnes coronales emphatiques / S D/ ne nécessite pas une rétraction de la Minf. Ce résultat peut également constituer un argument contre l'hypothèse selon laquelle les consonnes emphatiques sont accompagnées d'un arrondissement des lèvres (Jakobson, 1962). Cette déduction est basée sur les observations de (Mooshammer et al. 2007 ; Lee, 1995) qui montrent une position plus avancée de la Minf durant /ʃ/ comparée à /s/ attribuée à l'arrondissement des lèvres durant la première.

/d T D l r k g/ ont une position horizontale de la Minf qui reste statistiquement identique, bien que sa position verticale est significativement plus élevée durant /d T D/ comparée à /l r k g/. Ces résultats suggèrent également que le mouvement de translation horizontale (antérieur-postérieur) de la Minf peut être contrôlé indépendamment du mouvement de rotation.

/ħ/ présente la position de la Minf qui est la plus reculée comparée à toutes les autres consonnes (seule /ħ/ vs /q χ/ est non significative). /q χ/ ont une position horizontale de la Minf située entre les vélaires (différence non significative) et les pharyngales (différence non significative). La Minf durant /h/ est significativement plus reculée comparées à /s S t/, similaire à celle durant /b d T D l r k g/, mais plus avancée comparée à /q χ ħ/. Notons que /h/ présente un comportement très particulier. Sa hauteur varie significativement en fonction de la voyelle adjacente, alors que sa position horizontale reste quasi-stable (Table 3). Ce résultat, observé chez nos 3 locuteurs pris séparément, confirme que les mouvements de la Minf ne peuvent être réduits à une simple rotation.

3.3 Variations des positions de Minf durant /C/ en fonction des voyelles adjacentes

	b	s	S	t	d	T	D	l	r	k	g	q	χ	ħ	h
x-val	-0,2	-0,1	0,2	-0,2	0,3	0,2	0,1	0,7	0,2	1,1	1,1	0,3	1	1,1	0,1
a vs. i	ns	ns	ns	ns	ns	ns	Ns	ns	ns	ns	ns	ns	ns	ns	ns
y-val	-1,4	-0 ,1	-0,1	-0,1	-0,2	0,1	-0,1	-2,5	-2	-2,3	-2,5	-1,9	-1,9	-4,4	-3,7
a vs. i	ns	ns	ns	ns	ns	ns	Ns	ns	ns	ns	ns	ns	ns	***	***

TABLE 2 : Différences moyennes entre les valeurs de la position horizontale (x-val) et verticales (y-val) de /C/ dans aCa comparée à iCi (*, p<0.05, **, p<0.01, ***, p<0.001). (-) : différence négative

Nos données (TABLE 2) montrent que les variations de la position horizontale de la Minf durant chaque consonne dans aCa, comparée à sa correspondante dans iCi, sont moins importantes que les variations de sa position verticale. Par rapport à la dimension horizontale, les consonnes postérieures semblent subir un effet plus important du contexte vocalique comparées aux consonnes antérieures (différences moyennes plus faibles). Par rapport à la dimension verticale, les consonnes obstruantes coronales présentent les différences moyennes les plus réduites, alors que les pharyngales et les laryngales affichent les différences moyennes les plus importantes. La hauteur de la Minf est donc plus stable durant les obstruantes coronales, mais très influencée par le contexte vocalique durant les pharyngales et les laryngales, /b l r k g q x/ ont un comportent intermédiaire.

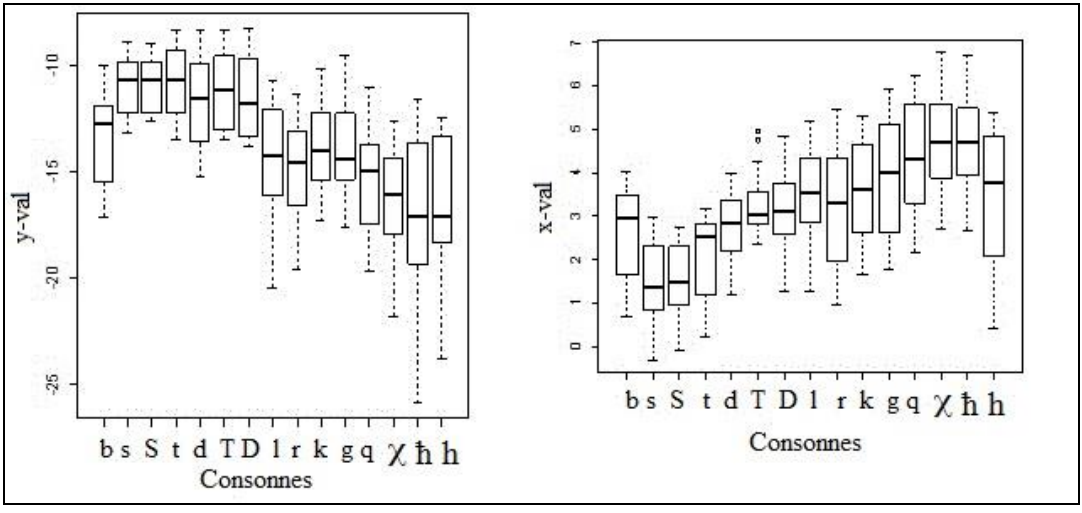


Figure 2 : Variation de la position verticale (à gauche) et horizontale (à droite) de la mâchoire inférieure (en mm) durant /b s S t d T D l r k g q χ ħ h/ de l'AM produites dans iCi et aCa. Chaque valeur correspond à la moyenne de 5 répétitions prononcées par 3 locuteurs dans les deux contextes.

	y-val			x-val		
	aCa	iCi	aCa + iCi	aCa	iCi	aCa + iCi
/b/	-14,1 (1,9)	-12,6 (2,4)	-13,3 (2,2)	2,2 (0,9)	2,4 (1,1)	2,6 (1,0)
/s/	-11,0 (1,3)	-10,9 (1,4)	-11,0 (1,4)	1,5 (0,9)	1,6 (1,1)	1,5 (1,0)
/S/	-10,9 (1,2)	-10,8 (1,4)	-10,9 (1,3)	1,7 (0,7)	1,4 (1,0)	1,7 (0,7)
/t/	-10,7 (1,6)	-10,7 (1,5)	-10,7 (1,5)	2,0 (0,9)	2,2 (1,0)	2,5 (1,0)
/d/	-11,8 (1,9)	-11,6 (2,3)	-11,7 (2,1)	2,9 (0,8)	2,6 (0,8)	2,9 (0,8)
/T/	-11,2 (1,6)	-11,2 (1,9)	-11,2 (1,7)	3,4 (0,8)	3,2 (0,6)	3,4 (0,8)
/D/	-11,5 (1,9)	-11,5 (1,8)	-11,5 (1,8)	3,3 (1,1)	3,1 (0,8)	3,3 (1,1)
/l/	-15,8 (2,6)	-13,4 (2,3)	-14,6 (2,7)	3,8 (1,0)	3,1 (1,0)	3,5 (1,0)
/r/	-15,8 (2,1)	-13,9 (2,2)	-14,8 (2,3)	3,3 (1,5)	3,1 (1,3)	3,2 (1,4)
/k/	-14,9 (1,6)	-12,7 (2,1)	-13,8 (2,2)	4,1 (1,1)	3,1 (0,9)	3,6 (1,1)
/g/	-15,2 (1,3)	-12,7 (2,0)	-13,9 (2,1)	4,3 (1,5)	3,2 (1,0)	3,8 (1,4)
/q/	-16,1 (1,8)	-14,3 (2,4)	-15,2 (2,3)	4,5 (1,5)	4,2 (1,0)	4,4 (1,3)
/x/	-17,2 (2,1)	-15,3 (2,2)	-16,2 (2,3)	5,2 (1,0)	4,2 (1,1)	4,7 (1,2)
/ħ/	-19,4 (3,3)	-15,0 (2,8)	-17,2 (3,7)	5,3 (1,2)	4,2 (0,9)	4,8 (1,2)
/h/	-18,5 (2,5)	-14,8 (2,3)	-16,7 (3,0)	3,4 (1,7)	3,3 (1,6)	3,3 (1,6)

TABLE 3 : Valeurs moyennes et (écart-type) de la position verticale (y-val) et horizontale (x-val) de la Minf (en mm) durant /b s S t d T D l r k g q χ ħ h/ prononcées 5 fois par 3 locuteurs dans aCa et iCi ainsi que dans ces deux contextes.

4 Conclusion

Cette étude a essayé d'expliquer les causes des variations des positions spatiales de la mâchoire inférieure durant plusieurs consonnes de l'arabe marocain ayant des lieux et modes d'articulation

différents et produites dans les contextes iCi et aCa. Nous avons montré que l'implication de la mâchoire inférieure est cruciale durant les obstruantes coronales /s S t T/. Le recul de la racine de la langue n'est pas nécessairement corrélé à la baisse de la Minf. De même que les consonnes apicales ne sont pas toujours associées à l'abaissement de la Minf. Nos données montrent également que la mâchoire inférieure ne semble pas être impliquée durant les laryngales et les pharyngales, ce qui est en accord avec les déductions de Goldstein (1994). Le dernier résultat majeur de notre étude montre que les mouvements verticaux (haut/bas) et horizontaux (avant-/arrière) de la Minf ne sont pas toujours corrélés ; les déplacements de la Minf durant la production des consonnes ne peuvent donc être réduits à un simple mouvement de rotation.

	b	t	d	T	D	s	S	l	r	k	g	q	χ	ħ	h
b		***	ns	**	*	***	***	ns	ns	Ns	ns	*	***	***	***
t			ns	ns	ns	ns	ns	***	***	***	***	***	***	***	***
d				ns	ns	ns	ns	***	***	**	**	***	***	***	***
T					ns	ns	ns	***	***	***	***	***	***	***	***
D						ns	ns	***	***	**	***	***	***	***	***
s							ns	***	***	***	***	***	***	***	***
S								***	***	***	***	***	***	***	***
l									ns	Ns	ns	ns	ns	***	**
r										Ns	ns	ns	ns	***	*
k											ns	ns	***	***	***
g												ns	**	***	***
q													ns	*	ns
χ														ns	ns
ħ															ns

TABLE 4 : Tests post-hoc TukeyHSD pour les positions verticales de la Minf (*, p<0.05, **, p<0.01, ***, p<0.001). La case noircie : Minf durant /C/ en colonne plus basse que /C/ en ligne.

	b	t	d	T	D	s	S	l	R	K	g	q	χ	ħ	h
b		ns	ns	ns	ns	**	*	Ns	Ns	*	**	***	***	***	ns
t			ns	**	*	ns	Ns	***	*	***	***	***	***	***	***
d				ns	ns	**	**	Ns	Ns	ns	*	***	***	***	ns
T						***	***	Ns	Ns	ns	ns	*	***	***	ns
D						***	***	Ns	Ns	ns	ns	***	***	***	ns
s							ns	***	***	***	***	***	***	***	***
S								***	***	***	***	***	***	***	***
l									Ns	ns	ns	ns	***	***	ns
r										ns	ns	**	***	***	ns
k											ns	ns	**	**	ns
g												ns	ns	*	ns
q													ns	ns	*
χ														ns	***
ħ															***

TABLE 5 : Tests post-hoc TukeyHSD pour les positions horizontales de la Minf (*, p<0.05, **, p<0.01, ***, p<0.001). La case noircie : Minf durant /C/ en colonne plus avancée que /C/ en ligne.

Références

DART S. (1991). "Articulatory and acoustic properties of apical and laminal articulations," *UCLA Working Papers in Phonetics* 79, 1–155.

ELGENDY A.M. (1999). Jaw contribution to the timing control of pharyngeal consonants production. Proc. of the *XIV*th *ICPhS*, San Francisco: 2415-2418.

GOLDSTEIN L. (1994). Possible articulatory bases for the class of guttural consonants. In P. Keating (ed.) *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III.* Cambridge: Cambridge University Press, 234-241.

HOOLE, P. & KÜHNERT, B. (1996). Tongue-jaw coordination in German vowel production. *Proc. of the 4th Speech Production Seminar*, Autrans, 97-100.

JAKOBSON, R. (1962). Muffxxama, the 'emphatic' phonemes in Arabic. In E. Pulgram (ed.). *Studies presented to Jushua Whatmough on his 60th Birthday*. The Hague: Mouton.105-115.

KEATING P., LINDBLOM B., LUBKER J., and KREIMAN J. (1994). "Variability in jaw height for segments in English and Swedish VCVs," *J. Phonetics* 22, 407–422.

LEE S.H. (1995). Orals, gutturals, and the jaw. In B. Connell and A. Arvaniti (eds.). *Phonology and phonetic evidence: Papers in Laboratory Phonology IV*. Cambridge: Cambridge University Press.

LINDBLOM B. (1983). Economy of speech gestures. In P.F. MacNeilage (ed.), *The Production of Speech*, 217-245. New York: Springer.

MOOSHAMMER C. HOOLE P. & GEUMANN A. (2007). Jaw and order, *Language Speech*50, 145-76.

MOOSHAMMER C., HOOLE P., GEUMANN A. (2006). Interarticulator cohesion within coronal consonant production. *Journal of the Acoustical Society of America*, 120, 1028-1039.

NOLAN, F. (1995). The role of the jaw active or passive?. In B. Connell & A. Arvaniti (eds.). *Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV*. Cambridge: Cambridge University Press.

RECASENS D. (2012). A study of jaw coarticulatory resistance and aggressiveness for Catalan consonants and vowels. *Journal of the Acoustical Society of America*, 132, 412-420.

SHADLE C. (1985). *The acoustics of Fricative Consonants*. PhD. Thesis MIT.

ZEROUAL C., HOOLE P., FUCHS S., ESLING J.H. (2007). EMA study of the coronal emphatic and non-emphatic plosive consonants of Moroccan Arabic. *Proc. of the XVI*th *ICPhS*, Saarbrücken: 397-340.

ZEROUAL C, HOOLE P, and ESLING J.H. (2011a). Contraintes articulatoires et acoustico-perceptives liées à la production de /k/ emphatisée en arabe marocain. In M. Embarki and C. Dodane (eds.), *La Coarticulation. Des indices à la représentation*, 227-240. Paris : L'Harmattan.

ZEROUAL C, ESLING J.H, HOOLE P, and RIDOUANE R. (2011). Ultrasound study of Moroccan Arabic labiovelarization. *Proc. of the XVII*th *ICPhS*, Hong Kong.

Étude transversale du rythme de l'anglais chez des apprenants francophones

Quentin Michardière[1], Anne Guyot-Talbot[1], Emmanuel Ferragne[1], François Pellegrino[2]
(1) CLILLAC-ARP EA 3967 – Université Paris Diderot
5 rue Thomas Mann, 75013 Paris, France
(2) Dynamique Du Langage UMR 5596 – CNRS / Université Lyon 2
14 avenue Berthelot, 69007 Lyon, France
emmanuel.ferragne@univ-paris-diderot.fr

RESUME

Dans cette étude, nous avons demandé à 3 groupes d'étudiants en anglais (niveaux L1, L2 et L3) de lire un dialogue en anglais afin d'évaluer la possibilité d'une amélioration de leur production sur le plan du rythme en fonction de leur niveau universitaire. Le dialogue a également été lu par des anglophones, et une traduction du dialogue en français a été enregistrée par des francophones dans le but d'établir un espace de référence pour l'anglais L1 et le français L1. Nous avons employé des mesures classiques du rythme s'appuyant sur la durée, et avons également exploré la pertinence de mesures basées sur l'enveloppe d'amplitude et le spectre de cette enveloppe. Nous constatons un manque de fiabilité des mesures classiques du rythme, et proposons quelques pistes.

ABSTRACT

A cross-sectional study of rhythm in French students of English.
In this study, 3 groups of French students of English (first, second and third year) were asked to read a dialogue in English in order to test whether their production improved in terms of rhythm depending on their academic level. The same dialogue was read by native English speakers, and a French translation by French speakers so that a reference space could be computed for English and French as first languages. Duration-based metrics were used. And we explored the relevance of parameters based on the amplitude envelope and the related spectrum. Most metrics turned out to be unreliable, and we suggest new possibilities.

MOTS-CLES : rythme de la parole, anglais langue étrangère.
KEYWORDS: speech rhythm, English as a foreign language.

1 Introduction

Il est établi que la perception d'un accent étranger est imputable aux caractéristiques segmentales, mais également suprasegmentales, de la parole (Trofimovitch & Isaacs, 2012). Concernant la sensation de rythme plus particulièrement, elle est d'autant plus cruciale dans l'acquisition de l'anglais par des apprenants francophones que les deux langues en question s'opposent par des caractéristiques rythmiques très différentes (Grabe & Low, 2002 ; Ramus et al., 1999). L'anglais est en effet un prototype de langue accentuelle, c'est-à-dire, une langue dans laquelle le signal contient une alternance de syllabes accentuées et de syllabes non accentuées, qui lui confère cette impression

d'irrégularité, rappelant subjectivement le code Morse. A l'inverse, le français dégage une impression auditive de régularité induite par une succession de syllabes tendant à présenter une durée et une intensité équivalente. On peut donc anticiper que si l'acquisition tardive de l'anglais par des apprenants francophones se caractérise, comme attendu, par des phénomènes de transfert, leur production en anglais devrait faire apparaître un rythme au moins partiellement typique du français.

1.1 Mesures de rythme de la parole

La grande majorité des mesures du rythme de la parole s'appuient sur la durée des segments, ou, plus précisément, des intervalles consonantiques ou vocaliques, chaque intervalle représentant une suite ininterrompue de chaque type, consonne ou voyelle. La plupart de ces indices peuvent être normalisés pour prendre en compte les variations de débit de parole. Certains consistent en une mesure globale sur un énoncé complet – comme ΔC et %V – alors que d'autres prennent en compte l'aspect séquentiel des intervalles. Dans cette dernière catégorie, on trouve les indices de type Pairwise Variability Index (PVI), qui calculent la différence de durée entre deux segments successifs du même type (consonantique ou vocalique), offrant ainsi, notamment, une meilleure modélisation des variations locales de débit.

En reprenant l'exemple de langues typiquement accentuelles et syllabiques, les études partent du constat que les langues accentuelles présentent généralement une tendance à la réduction des voyelles accompagnée d'une plus grande complexité phonotactique, alors que les langues syllabiques tendent à résister à la réduction vocalique et à avoir des structures syllabiques plus simples. La robustesse de ces mesures a été sévèrement remise en cause dans Arvaniti (2012), où l'auteure constate la grande sensibilité de ces indices de durée à des variations intra-langue : locuteur, méthode d'élicitation (spontané vs. lu), type de syllabe employé. Gut (2012) arrive à des conclusions similaires, mettant en évidence l'impossibilité de comparer les valeurs obtenues d'une étude à l'autre, et allant même jusqu'à suggérer l'abandon du concept de rythme dans la parole (voir aussi Nolan & Jeon, 2014).

Si la plupart des études se sont concentrées sur des mesures temporelles, il existe cependant des alternatives. Par exemple, Ferragne et Pellegrino (2008) ont appliqué le concept de PVI à l'intensité. Sur la base d'un corpus de 13 accents de l'anglais, cette mesure engendre des taux de classification automatique légèrement supérieurs à ceux induits par un PVI de durée. Tilsen et Arvaniti (2013) proposent des méthodes fondées sur la décomposition de l'enveloppe d'amplitude. La méthode que nous leur avons empruntée dans cet article consiste en une analyse spectrale de l'enveloppe d'amplitude du signal. Ce type d'analyse est encore à un stade quelque peu expérimental, en particulier parce qu'il ne reflète pas la différence entre les classes de rythme (Tilsen & Arvaniti, 2013). Cette approche vise à identifier des bandes de fréquences, l'une relativement basse reflétant l'influence d'un rythme « supra-syllabique » traduisant la présence d'accents, l'autre, relativement élevée correspondant au rythme syllabique. Tilsen et Arvaniti (2013) proposent le calcul du rapport de l'énergie contenu dans la bande 1,5 Hz-3,25 Hz avec l'énergie dans la bande 3,25-10 Hz, qu'ils appellent spectral band power ratio (SBPR). Le choix de ces valeurs est motivé par le fait que, pour la bande basse, censée capturer les intervalles entre accents, 1,5Hz permet des mesures sur une période de 667 ms ce qui est, pour les auteurs, tout à fait adapté comme durée maximale pour de la parole spontanée. La limite de 3,25Hz capture, quant à elle, une période de 308 ms, qui, pour les auteurs, correspond à la durée maximale d'une syllabe. Tilsen et Arvaniti (2013) reconnaissent volontiers une part d'arbitraire dans ces seuils, et il est probable qu'il faille les adapter pour de la parole d'apprenants. L'interprétation du SBPR est la suivante : des valeurs relativement élevées

traduisent une prédominance de périodicité accentuelle, alors que de faibles valeurs reflètent une périodicité plus syllabique dans l'enveloppe.

1.2 Rythme et langue seconde (L2)

L'existence de différents types de rythmes dans les langues – en particulier syllabique vs. accentuel – conduit tout naturellement à formuler l'hypothèse selon laquelle lorsque la L1 et la L2 d'un locuteur appartiennent à deux classes rythmiques différentes, un transfert négatif du rythme de la L1 vers la L2 s'opère. A l'inverse, dans le cas où la L1 et la L2 présentent un type de rythme identique, c'est un transfert positif qui se produit. Entre autres analyses, White et Mattys (2007) ont comparé des locuteurs natifs de l'anglais (langue accentuelle) s'exprimant en espagnol (langue syllabique) et des hispanophones s'exprimant en anglais. Ils utilisent les mesures de rythme classiques s'appuyant sur la durée des intervalles consonantiques ou vocaliques. Les auteurs précisent que les locuteurs, lorsqu'ils s'expriment dans leur L2, ont un accent étranger tout à fait évident. Les résultats varient en fonction de l'indice de rythme utilisé. Par exemple, ΔC ne fait apparaître aucune différence entre les anglophones s'exprimant en espagnol ou les hispanophones s'exprimant en anglais. En revanche, une mesure telle que VarcoV tend à révéler le schéma attendu : l'anglais L1 présente une valeur élevée, l'espagnol L1, une valeur basse, et ces deux langues produites comme L2 ont des valeurs intermédiaires.
Plus proches de notre étude, Tortel et Hirst (2010) ont analysé, toujours en s'appuyant sur les mesures classiques d'intervalles de durée, le rythme de 3 groupes de locuteurs : apprenants de l'anglais, intermédiaires et avancés, et locuteurs natifs. En combinant leurs 9 paramètres rythmiques, les auteurs, après avoir procédé à une analyse linéaire discriminante, obtiennent un taux de classification correcte de 69,5% entre les 3 groupes. Mais là encore, certains paramètres – en l'occurrence %V – présentent des valeurs non conformes aux attentes (relativement élevé pour l'anglais). Cette étude, comme la nôtre, est une des rares à proposer une évaluation de la parole en L2 faisant intervenir le niveau de compétence des apprenants.
La revue de question de Gut (2012) fait pleinement apparaître la fiabilité partielle des mesures de rythme les plus répandues, qui dépend grandement de la méthodologie employée. En particulier, les conclusions de cette revue mettent en évidence que les mesures de rythme appliquées à la comparaison entre des versions L1 et L2 d'une même langue, et a fortiori pour comparer des groupes d'apprenants de plusieurs niveaux, sont quasiment inutiles.
Dans ce contexte peu favorable, notre étude s'inscrit dans un cadre très exploratoire, et tente d'évaluer non seulement la possibilité d'utiliser les mesures classiques, mais également des indices moins connus basés sur l'enveloppe du signal.

2 Expérience

2.1 Corpus

Douze étudiants de chacun des trois niveaux de la licence d'anglais (L1, L2 et L3) à l'Université Paris Diderot ont lu un dialogue en anglais créé pour l'occasion, comportant 253 mots. Douze autres étudiants ont lu une traduction de ce dialogue en français (256 mots). Dix anglophones ont également été enregistrés, lisant le dialogue en anglais. Les participants francophones ont été recrutés sur la base du volontariat ; les anglophones ont enregistré le dialogue au début d'une autre étude pour laquelle ils percevaient une rémunération. Ces 58 participants ont été enregistrés dans une salle insonorisée avec un ordinateur portable par le biais d'un microphone USB Audio-Technica

AT 2020. Le signal a été numérisé au format PCM mono, 44,1 kHz, 16 bits. L'intégralité du dialogue était présentée sur un écran d'ordinateur avec le logiciel ROCme! (Ferragne et al. 2012), et les participants étaient invités à prendre connaissance du dialogue à leur rythme avant de l'enregistrer.

2.2 Analyses

Tous les fichiers ont été segmentés manuellement en phonèmes sous Praat. À partir de cette segmentation, les intervalles consonantiques et vocaliques – pour le calcul des mesures basées sur la durée – ont été automatiquement déterminés. Parmi les multiples mesures de rythme s'appuyant sur la durée disponibles dans la littérature, nous nous sommes contentés d'utiliser les 6 mesures de l'étude d'Arvaniti (2012) : 1) ΔC : l'écart-type de la durée des intervalles consonantiques dans un énoncé 2) %V : le pourcentage de durée des intervalles vocaliques d'un énoncé, 3) rPVIC : la moyenne, sur un énoncé, de la différence absolue de durée entre deux intervalles consonantiques successifs, 4) nPVIV : la moyenne, sur un énoncé, de la différence absolue de durée entre deux intervalles vocaliques successifs, la différence de chaque paire étant normalisée par la somme des durées des deux intervalles, divisée par 2, 5) VarcoC : le coefficient de variation de la durée des intervalles consonantiques sur un énoncé, et 6) VarcoV : le coefficient de variation de la durée des intervalles vocaliques sur un énoncé. Les premières analyses montraient une forte variation intra-groupe et intra-individuelle ; nous avons donc émis l'hypothèse que cela pouvait provenir de la nature de notre dialogue. En effet, la plupart des phrases étant particulièrement courtes et prononcées avec emphase (ex. *How are you*), nous avons craint qu'elles donnent lieu à des valeurs déviantes. Nous avons donc restreint notre analyse à quatre phrases, parmi les plus longues, sur lesquelles on pouvait anticiper que le calcul du rythme soit plus fiable. Ces phrases sont : P1) Do you know what the Dalai Lama said at Subway? P2) Do you really need to explain the jokes you are told? P3) It's because sandwiches are very good linguistic material, P4) Because imagine you suddenly turn back into your former self in my stomach. En français : P1) Tu sais ce que dit le Dalai Lama au Subway ? P2) Tu as besoin d'expliquer les blagues qu'on te raconte ? P3) C'est parce que les sandwichs sont de bons exemples en linguistique, P4) Parce qu'imagine, d'un coup, tu reprends ta forme humaine dans mon estomac. Sauf indication contraires, les analyses s'appuient sur la moyenne de ces 4 phrases pour chaque locuteur.

Pour les métriques s'appuyant sur l'enveloppe spectrale, notre méthode s'inspire très largement de Tilsen et Arvaniti (2013). Un premier filtrage entre 400 et 4000 Hertz est effectué, dont le but est d'atténuer l'impact de F0 dans le calcul final (coupure à 400 Hz), ainsi que d'éviter que des maxima de l'enveloppe soient dus à des sifflantes ou des occlusives. Le signal est ensuite soumis à un filtre passe-bas à 10 Hz, permettant d'extraire ce que nous allons considérer comme l'enveloppe d'amplitude (FIGURE 3). Une transformée de Fourier est ensuite appliquée à cette enveloppe, qui nous permet d'obtenir un spectre de puissance. Nous calculons alors deux indices : le centre de gravité du spectre, et le rapport de l'énergie contenue dans la bande 1,5 Hz-3,25 Hz et dans la bande 3,25-10 Hz.

2.3 Résultats et discussion

2.3.1 Mesures de référence du rythme

Dans un premier temps, les données sont représentées dans les 3 espaces de référence dans la FIGURE 1 : %V/ΔC, nPVIV/rPVIC et VarcoV/VarcoC.

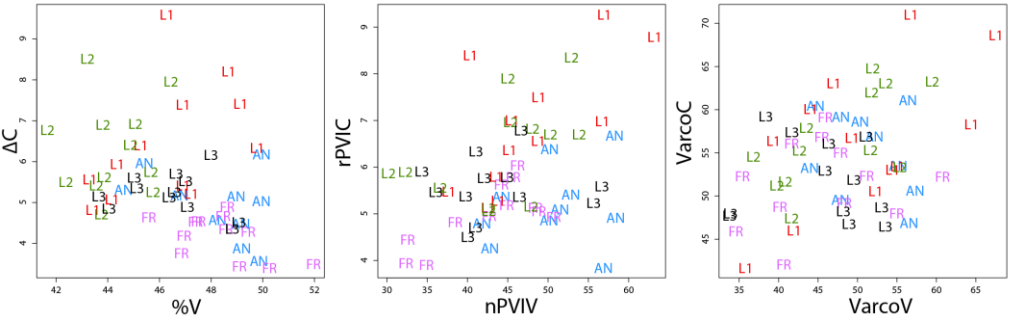


FIGURE 1 : Dispersion des locuteurs dans les espaces de rythme de référence

Dans ces 3 espaces, il paraît difficile d'isoler chaque groupe. Cette impression se confirme dès lors qu'on tente d'apprendre des modèles d'analyse linéaire discriminante (LDA) pour espace de référence. Le modèle entrainé sur les espaces %V/ΔC, nPVIV/rPVIC et VarcoV/VarcoC donnent respectivement des taux de classification correcte de 52%, 38% et 59%. Il est donc impossible de classifier les locuteurs sur la base de ces mesures (on note cependant que 10 des 12 locuteurs FR sont correctement classés dans le premier espace). Ce qui est encore plus surprenant, c'est que les productions des locuteurs AN et FR, pourtant supposés représenter deux langues aux caractéristiques rythmiques radicalement opposées, ne semblent pas former des groupes très distincts. Si on entraine des modèles de LDA seulement avec les locuteurs AN et FR, les taux de classification restent en effet décevants : 68%, 68% et 59% pour chacun des espaces décrits plus haut. Lorsqu'on prend en compte tous les paramètres avec un modèle incluant les 5 classes, on atteint 55% de classification correcte.

En regardant la FIGURE 1, on peut noter que le schéma d'un transfert négatif, selon lequel les apprenants s'exprimant en anglais (groupes L1, L2 et L3) présenteraient des valeurs intermédiaires entre leur langue maternelle et leur langue seconde, ne s'applique pas. Ceci est particulièrement visible dans le plan %V/ΔC, où les échantillons de langue maternelle (AN et FR) se regroupent dans le coin inférieur droit. Dans ce même espace, on remarque une grande variation de ΔC pour les L1 et L2, mais une plus grande homogénéité pour les L3.

2.3.2 Effet du débit

Il a souvent été remarqué que ces mesures étaient particulièrement sensibles au débit de parole (Dellwo, 2006). On relève en effet, comme on pouvait s'y attendre, des corrélations élevées entre débit de parole (débit vocalique) et certains indices non normalisés, en particulier ΔC (r=-0,80) et rPVIC (r=-0,81). La FIGURE 2 montre la variation de débit vocalique en fonction du groupe.

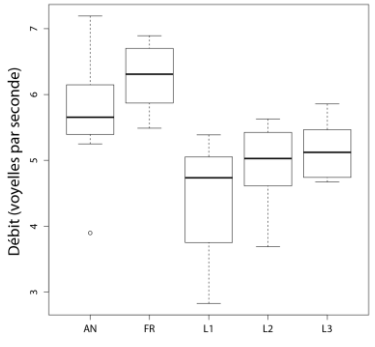


FIGURE 2 : Débit vocalique par groupe

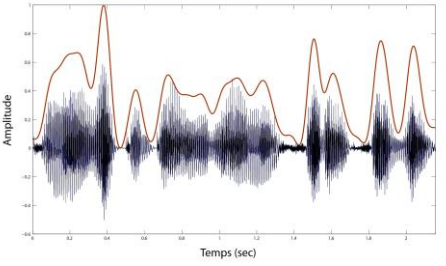


FIGURE 3 : Enveloppe d'amplitude pour la phrase P1 du locuteur BENO (anglophone)

Une analyse de la variance à un facteur montre un effet significatif du groupe sur le débit vocalique ($F(4,53)=13,95$; $p<0,001$). Des comparaisons post-hoc font apparaître des différences significatives entre les débits du groupe FR et celui de tous les groupes d'apprenants, et entre le groupe AN et le groupe L1. On note par ailleurs, visuellement, une réduction de la variabilité intra-groupe au fur et à mesure de l'avancée dans le cursus (L1, L2, L3). L'origine de ce déséquilibre entre groupes de natifs et groupes d'apprenants met en avant un débit de parole plus lent chez les apprenants, mais on peut également faire l'hypothèse que la lecture est plus difficile dans une langue étrangère. Trouvain et Möbius (2014) analysent le débit de parole de locuteurs germanophones s'exprimant dans leur L1 et en français, et de francophones s'exprimant dans leur L1 et en allemand. Leurs résultats confirment ce qui semble être un fait établi depuis longtemps dans la littérature : les productions en L2 affichent un débit plus lent. Le niveau de compétence en L2 de leurs locuteurs est également pris en compte, et les auteurs montrent une augmentation du débit avec le niveau des locuteurs, ce qui apparaît visuellement comme une tendance dans la FIGURE 2.

2.3.3 Enveloppe d'amplitude

Les rapports d'énergie entre les deux bandes de fréquences décrites plus haut (spectral band power ratio: SBPR) et le centre de gravité du spectre (COG) sont deux variables très corrélées : r=-0,89. Intuitivement, ces deux variables doivent dépendre en partie du débit de parole ; c'est en effet le cas : les corrélations avec le débit vocalique sont de -0,66 et 0,72 pour SBPR et COG. Une bonne estimation des caractéristiques rythmiques des langues après neutralisation du débit paraît dès lors compromise.

Nous avons également exploité la comparaison de l'enveloppe d'amplitude des phrases dans le domaine temporel. Cette enveloppe a été estimée par la méthode de filtrages successifs de Tilsen et Arvaniti (2013) décrite plus haut. L'alternance de syllabes accentuées et inaccentuées caractéristique de l'anglais devrait être matérialisée par une enveloppe présentant une succession de bosses d'amplitude variable. À l'inverse, la tendance du français à présenter une succession de syllabes de saillance équivalente devrait conduire à observer une enveloppe marquée par des bosses de taille équivalente. Les enveloppes de toutes les phrases ont été normalisées en amplitude, et chaque phrase a été comparée aux productions de cette même phrase par les autres locuteurs par le biais de la déformation temporelle dynamique (*dynamic time warping*, DTW) afin de tenter d'effacer les

variations de débit. On obtient ainsi une matrice de distances par phrase (P1, P2, P3 et P4) entre les 58 locuteurs. Puis on calcule la moyenne des 4 matrices de distances. On reconstruit ensuite un espace à 2 dimensions par le biais du *multidimensional scaling* (MDS), et les coordonnées des locuteurs dans l'espace MDS sont utilisées pour construire un modèle d'analyse linéaire discriminante. Le modèle d'analyse discriminante entraîné simplement sur les catégories natives (FR et AN) permet une classification correcte de 21 des 22 locuteurs (95% - un locuteur FR est mal classé). Entraîné sur toutes les classes ; le modèle retombe à 57% de classification correcte.

2.4 Conclusion

Les mesures du rythme aujourd'hui classiques – %V/ΔC, nPVIV/rPVIC et VarcoV/VarcoC – ont suscité beaucoup d'enthousiasme dans les années 1990 et 2000 (Dellwo, 2006 ; Grabe & Low, 2002 ; Ramus et al., 1999). Mais contrairement à d'autres paramètres phonétiques dans l'étude de la parole (formants, F0), la notion très impressionniste de rythme dans les langues, et l'opportunité de mesurer le phénomène de façon objective, continue de donner lieu à controverse. Nos résultats viennent appuyer la tendance actuelle, qui consiste à déplorer le manque de robustesse des mesures courantes (Gut, 2012 ; Arvaniti, 2012). En particulier, l'impossibilité dans notre étude de séparer les locuteurs de l'anglais des locuteurs du français (deux langues prétendument très distantes sur le plan du rythme) à partir des mesures classiques laisse perplexe. On peut envisager plusieurs explications partielles. Il est possible que les phrases choisies ne soient pas typiquement représentatives des classes de rythme des langues en question. Par exemple, la phrase P3 en français contient des syllabes dont la complexité s'apparente à celle des langues accentuelles. Cette particularité rappelle l'expérience d'Arvaniti (2012), dans laquelle les phrases à lire étaient choisies pour être plus ou moins typiquement accentuelles ou syllabiques, que la langue soit elle-même accentuelle ou syllabique. Ses résultats montrent que ce facteur module fortement la distance entre langues syllabiques et accentuelles. On peut également envisager que la méthode d'élicitation (parole lue) ait tendance à gommer les caractéristiques rythmiques des langues, en particulier si on prend en compte la difficulté supplémentaire, évidente dans les enregistrements, que représente le fait de lire. Arvaniti (2012) montre un effet de la méthode d'élicitation, mais qui n'affecte pas toutes les mesures de la même manière, et surtout, les affecte différemment en fonction de la langue.

Les mesures s'appuyant sur l'analyse spectrale de l'enveloppe d'amplitude (SBPR et COG) ne permettent pas de séparer les différentes classes de locuteurs, à l'exception d'une distinction entre locuteurs natifs (AN et FR) et apprenants, reflétant des différences de débit de parole. La détermination précise (en fonction du débit) de la coupure entre les deux bandes de fréquences utilisées pour calculer le SBPR semble nécessaire pour utiliser cette technique, et constitue donc un prolongement possible de notre étude. En revanche, la distance DTW entre l'enveloppe d'amplitude de chaque phrase permet, après MDS, de positionner les locuteurs dans un nouvel espace qui, puisqu'il permet une bonne discrimination FR vs. AN, reflète probablement des propriétés rythmiques. Une exploration plus précise de cette méthode constitue un développement intéressant de la présente étude. Concernant la possibilité de mesurer le rythme des productions des apprenants, notre méthodologie et nos mesures ne permettent pas de distinguer les 3 groupes. Nous n'avons, bien sûr, aucune certitude quant à l'amélioration effective du niveau d'anglais oral des apprenants entre la première et la troisième année à l'université sur la lecture du dialogue que nous leur avons proposé. Une tâche de perception impliquant des locuteurs natifs pourrait apporter un début de réponse sur ce point. On note cependant une tendance à l'homogénéisation des groupes en fonction de l'avancée dans le cursus universitaire (FIGURE 1 et FIGURE 2), et notre expérience d'enseignants nous conduit à penser que la prononciation des étudiants s'améliore effectivement sur les 3 années concernées.

L'étude phonétique du rythme dans les langues repose depuis une vingtaine d'année sur les mesures temporelles que nous avons employées, mais leur robustesse tout relative, conduit, comme le font Nolan et Jeon (2014) à remettre en question l'axiome de départ selon lequel les langues sont rythmiques et qu'il existe une méthode indépendante de la langue pour mesurer ce rythme. Nolan et Jeon (2014) développent l'idée que le rythme de la parole n'est qu'une analogie (avec le rythme de la musique par exemple) et qu'en fonction de la manière dont on peut synchroniser la structure linguistique, intrinsèquement arythmique, d'une langue avec une prétendue horloge externe, cela peut générer une impression de rythme plus ou moins marquée. Nos résultats, associés aux publications récentes dans le domaine (Arvaniti, 2012 ; Gut, 2012 ; Nolan & Jeon, 2014), encouragent à envisager la question du rythme différemment, et à s'affranchir des mesures classiques s'appuyant sur la durée.

Remerciements

Cette étude a bénéficié du soutien de l'IUF (E. Ferragne) et de l'Idex USPC (projet SOPHOCLE).

Références

ARVANITI A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, 40(3), 351-373.

DELLWO V. (2006). Rhythm and speech rate: A variation coefficient forΔ C. *Language and language-processing*, 231-241.

FERRAGNE E., FLAVIER S., FRESSARD C. (2012). ROCme! (Version 2.0) [Logiciel]. Consulté le 10 février 2016. *Téléchargeable à l'adresse :* www.ddl.ish-lyon.cnrs.fr/rocme

FERRAGNE E., PELLEGRINO F. (2008). Le rythme dans les dialectes de l'anglais: une affaire d'intensité? Actes de *Journées d'Etude de la Parole*, Avignon, 9-13.

GRABE E., LOW E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in laboratory phonology*, 7(515-546).

GUT U. (2012). Rhythm in L2 speech. *Speech and Language Technology*. 14/15, 83-94.

NOLAN F., JEON H. S. (2014). Speech rhythm: a metaphor?. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol 369(n°1658).

RAMUS F., NESPOR M., MEHLER J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73(3), 265-292.

TILSEN S., ARVANITI A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages. *The Journal of the Acoustical Society of America*, 134(1), 628-639.

TORTEL A., HIRST D. (2010). Rhythm metrics and the production of English L1/L2. *Proceedings of Speech Prosody*.

TROFIMOVICH P, ISAACS T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, 15, 905–916

TROUVAIN J., MÖBIUS B. (2014). Sources of variation of articulation rate in native and non-native speech: comparisons of French and German. *Proceedings of Speech Prosody* (SP7), 275-279.

WHITE L., MATTYS S. L. (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics*, 35(4), 501-522.

Exploration de paramètres acoustiques dérivés de GMMs pour l'adaptation non supervisée de modèles acoustiques à base de réseaux de neurones profonds

Natalia Tomashenko[1, 2, 3] Yuri Khokhlov[3] Anthony Larcher[2] Yannick Estève[2]
(1) ITMO University, Saint-Pétersbourg, Russie
(2) LIUM, Le Mans, France
(3) STC-innovations Ltd, Saint-Pétersbourg, Russie
prenom.nom@univ-lemans.fr
khokhlov@speechpro.com

RÉSUMÉ

L'étude présentée dans cet article améliore une méthode récemment proposée pour l'adaptation de modèles acoustiques markoviens couplés à un réseau de neurones profond (DNN-HMM). Cette méthode d'adaptation utilise des paramètres acoustiques dérivés de mixtures de modèles Gaussiens (*GMM-derived features*, *GMMD*). L'amélioration provient de l'emploi de scores et de mesures de confiance calculés à partir de graphes construits dans le cadre d'un algorithme d'adaptation conventionnel dit de *maximum a posteriori* (MAP). Une version modifiée de l'adaptation MAP est appliquée sur le modèle GMM auxiliaire utilisé dans une procédure d'apprentissage adaptatif au locuteur (*speaker adaptative training*, SAT) lors de l'apprentissage du DNN. Des expériences menées sur le corpus Wall Street Journal (WSJ0) montrent que la technique d'adaptation non supervisée proposée dans cet article permet une réduction relative de $8,4\%$ du taux d'erreurs sur les mots (WER), par rapport aux résultats obtenus avec des modèles DNN-HMM indépendants du locuteur utilisant des paramètres acoustiques plus conventionnels.

ABSTRACT

Exploring GMM-derived features for unsupervised adaptation of deep neural network acoustic models

In this paper we investigate GMM-derived features recently introduced for adaptation of context-dependent deep neural network HMM (CD-DNN-HMM) acoustic models. We present an initial attempt of improving of the previously proposed adaptation algorithm by applying lattice scores and by using confidence measures in the traditional maximum a posteriori (MAP) adaptation algorithm. Modified MAP adaptation is performed for the auxiliary GMM model used in a speaker adaptive training (SAT) procedure for a DNN. Experimental results on the Wall Street Journal (WSJ0) corpus show that the proposed adaptation technique can provide, on average, an $8,4\%$ relative word error rate (WER) reduction under an unsupervised adaptation setup, compared to speaker independent DNN-HMM systems built on conventional features.

MOTS-CLÉS : adaptation au locuteur, réseaux de neurones profonds, MAP, CD-DNN-HMM, paramètres acoustiques dérivés de GMM (GMMD), apprentissage adaptatif au locuteur (SAT).

KEYWORDS : speaker adaptation, deep neural networks (DNN), MAP, CD-DNN-HMM, GMM-derived parameters (GMMD), speaker adaptive training (SAT).

1 Introduction

Aujourd'hui, les réseaux de neurones profonds (DNNs) ont détrôné les modèles GMM-HMMs dans la plupart des systèmes état-de-l'art de reconnaissance de la parole (RAP), depuis qu'il a été montré que les modèles DNN-HMM obtiennent de meilleurs résultats dans plusieurs tâches de RAP (Hinton *et al.*, 2012). De nombeux algorithmes d'adaptation ont été développés pour les modèles GMM-HMM (Gales, 1998 ; Gauvain & Lee, 1994) mais ne peuvent pas être facilement appliqués aux DNNs en raison des différentes natures de ces systèmes. Les GMMs sont des modèles génératifs appris par maximisation de la vraisemblance sur les données d'apprentissage alors que les DNNs sont des modèles discriminant, dont les paramètres sont estimés pour minimiser les erreurs de classification. Puisque l'estimation des paramètres des DNNs est basée sur un critère discriminant, elle est plus sensible aux erreurs d'étiquettes et moins pertinente pour une adaptation non supervisée.

Plusieurs méthodes d'adaptation ont récemment été proposées pour les DNNs, et quelques unes (Rath *et al.*, 2013 ; Seide *et al.*, 2011 ; Lei *et al.*, 2013 ; Tomashenko & Khokhlov, 2014, 2015 ; Liu & Sim, 2014 ; Kanagawa *et al.*, 2015) tirent avantage de l'adaptabilité des GMMs. Cependant il n'existe pas de méthode universelle pour transférer de manière efficiente les algorithmes d'adaptation du paradigme gaussien vers celui des DNNs. L'objectif de cette étude est de faire un pas dans cette direction en utilisant des paramètres acoustiques dérivés de GMM pour estimer des modèles DNN.

La plupart des méthodes existantes pour adapter les modèles DNN peuvent être classées en quatre catégories : (1) les transformations linéaires, (2) les techniques de régularisation, (3) les paramètres auxiliaires, (4) les combinaisons de GMM et DNN. **La transformation linéaire** est une des approches les plus populaires pour l'adaptation des réseaux de neurones (ANN). Elle peut être appliquée à différents niveaux d'un système ANN-HMM : sur les paramètres d'entrée, avec une transformation linéaire (*linear input network transformation*, LIN) (Gemello *et al.*, 2006 ; Neto *et al.*, 1995 ; Li & Sim, 2010 ; Trmal *et al.*, 2010) ou avec une régression linéaire discriminante sur l'espace des paramètres (*feature-space discriminative linear regression, fDLR)* (Seide *et al.*, 2011 ; Yao *et al.*, 2012) ; sur les activations des couches cachées (*linear hidden network transformation*, LHN) (Gemello *et al.*, 2006 ; Neto *et al.*, 1995) ; sur la couche softmax, comme avec LON (Li & Sim, 2010) ou avec une régression linéaire discriminante sur les paramètres de sortie (*output-feature discriminative linear regression*) (Yao *et al.*, 2012). L'adaptation de modèles acoustiques hybrides par partage de probabilités *a posteriori* (Stadermann & Rigoll, 2005) peut également être considérée comme une transformation linéaire de ces probabilités. Enfin, les auteurs de (Dupont & Cheboub, 2000) décrivent une méthode qui repose sur une transformation linéaire dans l'espace des paramètres et sur une analyse en composantes principales. La seconde catégorie de méthodes d'adaptation consiste à réestimer complètement le réseau ou seulement une partie à l'aide de **techniques de régularisation** spécifiques pour améliorer la généralisation, comme la régularisation *L2-prior* (Liao, 2013), la régularisation par divergence de Kullback-Leibler (Yu *et al.*, 2013), ou l'apprentissage conservatif (Albesano *et al.*, 2006). Dans (Stadermann & Rigoll, 2005), seul un sous-ensemble des neurones cachés avec une variance maximale (calculée sur les données d'adaptation) est réestimé. Le nombre de paramètres spécifiques au locuteur est réduit dans (Xue *et al.*, 2014) à travers une factorisation basée sur une décomposition en valeurs singulières, et un apprentissage adaptatif régularisé d'un sous-ensemble de paramètres de DNN est exploré dans (Ochiai *et al.*, 2014).

L'utilisation de paramètres auxiliaires est une autre approche pour laquelle les vecteurs de paramètres acoustiques sont augmentés de paramètres additionnels spécifiques au locuteur ou au canal. Ces paramètres sont calculés pour chaque locuteur ou pour chaque phrase, à la fois pendant l'apprentissage

et pendant le décodage. Les i-vecteurs sont un exemple de paramètres auxiliaires efficaces (Senior & Lopez-Moreno, 2014 ; Saon *et al.*, 2013) ; il a été montré qu'ils étaient complémentaires avec une adaptation fMLLR. L'adaptation par codes de locuteurs (Abdel-Hamid & Jiang, 2013) et l'adaptation factorisée (Li *et al.*, 2014) sont des méthodes alternatives qui prennent en compte les facteurs sous-jacents qui contribuent à dégrader le signal de parole.

Le moyen le plus courant pour **combiner les modèles GMM et DNN** à des fins d'adaptation consiste à utiliser comme entrées, lors de l'apprentissage du DNN, des paramètres GMM ayant été adaptés, par exemple par fMLLR (Rath *et al.*, 2013 ; Seide *et al.*, 2011 ; Kanagawa *et al.*, 2015). Dans (Lei *et al.*, 2013), les scores de vraisemblance des modèles DNN et GMM, adaptés au niveau de l'espace des paramètres en utilisant la même transformation fMLLR, sont combinés au niveau des états pendant le décodage. Les auteurs de (Liu & Sim, 2014) proposent de combiner les modèles GMM et DNN en utilisant une approche par régression des poids variant dans le temps (*temporally varying weight regression, TVWR).* Dans la méthode d'apprentissage de la couche *bottleneck* dépendant du locuteur décrite dans (Doddipatla *et al.*, 2014), les paramètres du *bottleneck* normalisé par locuteur sont calculés et utilisés pour l'apprentissage de modèle GMM-HMM. Dans les systèmes de type *tandem*, les paramètres dérivés des réseaux de neurones sont également utilisés pour l'apprentissage des GMM (Ellis & Reyes-Gomez, 2001).

Une autre approche d'adaptation basée sur les modèles s'appuie sur l'analyse des contributions des neurones cachés spécifiques au locuteur (*learning speaker-specific hidden unit contributions*, LHUC). Dans (Siniscalchi *et al.*, 2013), la forme de la fonction d'activation est modifiée pour mieux correspondre aux caractéristiques liées au locuteur.

Dans le passé, il a été montré dans (Pinto & Hermansky, 2008) que les log-vraisemblances de GMM peuvent être utilisées avec succès comme paramètres pour estimer un système de reconnaissance de phonèmes construit sur la base d'un MLP (*multi-layer perceptron*). Dans notre étude, nous explorons une nouvelle approche de type *speaker adaptative training* (SAT) appliquée sur les DNN et basée sur l'utilisation de paramètres dérivés de GMM (GMMD) en entrée du réseau de neurones (Tomashenko & Khokhlov, 2014, 2015). Notre approche s'appuie également sur l'utilisation de techniques d'adaptation propres aux GMMs, appliquées aux GMMD.

Dans cet article, nous présentons une première tentative d'amélioration de cette approche en utilisant des graphes de reconnaissance durant l'adaptation MAP. La suite de l'article est organisée comme suit. Dans la Section 2, l'approche SAT pour les DNN-HMM basés sur les paramètres dérivés de GMM est introduite. La Section 3 décrit l'algorithme d'adaptation MAP utilisant des scores de graphes. Les résultats expérimentaux sont donnés en Section 4. Enfin, une conclusion est fournie en Section 5.

2 Apprentissage adaptatif au locuteur pour les modèles DNN-HMM basés sur des paramètres dérivés de GMMs

La construction de paramètres dérivés de GMMs pour l'adaptation de DNNs a été proposée dans (Tomashenko & Khokhlov, 2014, 2015), où il a été montré que ces paramètres rendent possible l'utilisation de techniques d'adaptation de HMM-GMMs dans le paradigme DNN, par exemple au travers d'une adaptation MAP ou fMLLR . Nos paramètres sont obtenus comme suit (voir Figure 1) : tout d'abord, 13 coefficients cepstraux de fréquence Mel (MFCC) et leurs coefficients Δ et $\Delta\Delta$ (au total 39 coefficients) sont extraits, et une normalisation par la moyenne cepstrale (CMN) est appliquée par

locuteur. Ensuite, un modèle GMM-HMM auxiliaire monophone indépendant du locuteur est utilisé pour transformer les vecteurs de paramètres cepstraux en vecteurs de vraisemblances. Au cours de cette étape, une adaptation au locuteur du modèle GMM-HMM est réalisée pour chaque locuteur du corpus d'apprentissage et le nouveau modèle SA GMM-HMM (SA : adapté au locuteur) créé est utilisé pour obtenir les paramètres dérivés de GMMs adaptés au locuteur. Dans le modèle GMM-HMM

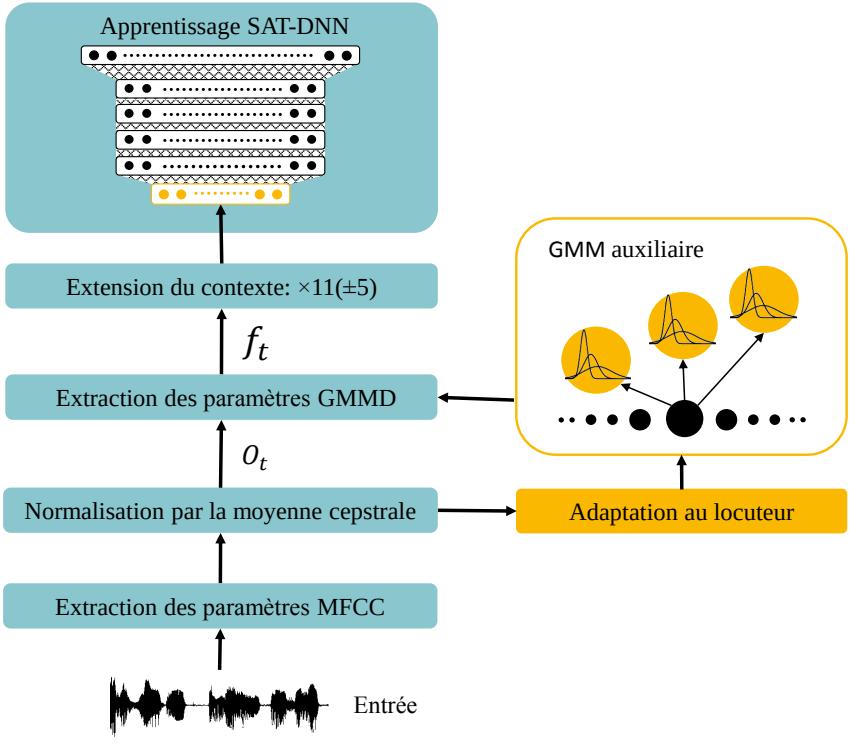


FIGURE 1 – Utilisation de paramètres dérivés de GMMs adaptés pour une apprentissage SAT de DNN-HMM.

auxiliaire, chaque phonème est modélisé indépendamment du contexte avec un modèle de Markov à trois états. Pour un vecteur de paramètres MFCC donné, un nouveau vecteur de paramètres dérivés de GMMs est obtenu en calculant les log-vraisemblances de tous les états des modèles monophones GMM-HMM. En supposant que o_t est un paramètre acoustique à l'instant t, alors le vecteur f_t de paramètres dérivés de GMMs est calculé comme suit :

$$f_t = [p_t^1, \ldots, p_t^n] \tag{1}$$

où n est le nombre d'états du modèle monophone auxiliaire GMM-HMM, et :

$$p_t^i = \log\left(P(o_t \mid s_t = i)\right) \tag{2}$$

est la log-vraisemblance estimée par le GMM-HMM. Ici, s_t est l'indice de l'état du HMM à t. Dans notre cas, n est égal à 132 ($= 39 \times 3 + 3 \times 5$) : 39 HMM à trois états (un HMM par phonème), un

modèle de silence modélisé par un HMM à cinq états, et deux modèles HMM de bruits (un parole et un non-parole) à 5 états. Nous obtenons ainsi, pour chaque trame, un vecteur à 132 dimensions de paramètres dérivés de GMMs.

Enfin, ces paramètres sont concaténés sur une fenêtre temporelle de 11 trames (± 5 autour de la trame visée). Ces 1452 (11×132) paramètres par trame, appelés GMMD, sont utilisés par la suite comme entrée pour l'apprentissage des DNN .

3 Utilisation de scores de graphes de décodage pour l'adaptation MAP

L'utilisation d'informations et de mesures de confiance provenant de graphes est une méthode connue pour améliorer les performances d'une adaptation non supervisée (Uebel & Woodland, 2001 ; Gollan & Bacchiani, 2008). Dans cette étude, nous utilisons l'algorithme d'adaptation MAP pour adapter des modèles GMM-HMM indépendants du locuteur (Gauvain & Lee, 1994). L'adaptation au locuteur d'un modèle DNN-HMM construit à partir de paramètres GMMD est réalisé au moyen d'une adaptation MAP des modèles monophones GMM-HMM auxiliaires, qui ont été utilisés pour calculer les paramètres GMMD.

Nous avons modifié l'algorithme classique d'adaptation MAP en utilisant des graphes au lieu d'un alignement sur la meilleure hypothèse, comme expliqué ci-dessous. Notons m l'indice d'une Gaussienne dans un modèle acoustique indépendant du locuteur, et μ_m la moyenne de cette Gaussienne. Alors l'estimation MAP du vecteur des moyennes est :

$$\widehat{\mu}_m = \frac{\tau \mu_m + \sum_t \gamma_m(t) p_s(t) o_t}{\tau + \sum_t \gamma_m(t) p_s(t)} \tag{3}$$

où τ est le paramètre qui contrôle l'équilibre entre l'estimation de la moyenne par maximum de vraisemblance et sa valeur *a priori*, $\gamma_m(t)$ est la probabilité *a posteriori* du composant gaussien m à l'instant t, et $p_s(t)$ est la mesure de confiance pour l'état s à l'instant t obtenue en calculant les probabilités *a posteriori* des arcs dans le graphe des états de la première passe de décodage.

L'algorithme *forward-backward* est utilisé pour calculer ces probabilités *a posteriori* à partir du graphe. Notons que lorsque $p_s(t) = 1$ pour tous les états, la formule (3) represente l'adaptation MAP classique.

En plus de cette pondération au niveau de la trame, nous appliquons une stratégie de sélection basée sur la mesure de confiance en n'utilisant dans la formule (3) que les observations dont les scores de confiance dépassent un seuil fixé *a priori*.

Pour l'adaptation des modèles acoustiques DNN, l'adaptation MAP est d'abord appliquée sur les modèles monophones GMM-HMM indépendant du locuteur pour créer un modèle adapté au locuteur SA GMM-HMM, comme nous venons de le décrire. Ensuite, au moment de reconnaissance de la parole, les GMMD sont calculés à partir de ce modèle SA GMM-HMM. L'approche proposée peut être considérée comme une technique de transformation dans l'espace des paramètres, puisque les DNN-HMMs sont appris sur des paramètres dérivés de GMMs.

4 Résultats expérimentaux

Dans ce travail préliminaire, nous présentons les résultats obtenus en suivant le protocole standard WSJ0 **si_et_20** (Paul & Baker, 1992) qui comporte 333 phrases lues (5645 mots) par 8 locuteurs. Nous utilisons un modèle de langage trigramme contenant 20000 mots (*open NVP LM*). Le taux de mots hors vocabulaire est de $1,5\%$. Ce modèle de langage est réduit selon le procédé décrit dans la recette KALDI pour WSJ (Povey *et al.*, 2011) avec un seuil à 10^{-7}.

Les modèles acoustiques sont estimés sur 7138 phrases prononcées par 83 locuteurs extraits des données d'apprentissage standard SI-84, soit environ 13 heures de parole et 2 heures de silence enregistrées en 16kHz. Le jeu de 39 phonèmes est complété par un modèle de silence et deux modèles de bruit. Les modèles acoustiques sont appris avec la plateforme KALDI (Povey *et al.*, 2011) en suivant la recette WSJ fournie, excepté pour l'extraction des paramètres GMMD et l'adaptation du modèle. Notre système de référence utilise 13 paramètres MFCC avec leurs dérivées premières et secondes. Ces 39 paramètres, utilisés avec leur contexte de 11 trames (5 avant et après) sont comparés aux paramètres GMMD proposés.

Trois réseaux de neurones profonds (DNN) sont estimés : un modèle de référence, indépendant du locuteur (SI) utilisant 11×39MFCC ; deux modèles utilisant les paramètres GMMD : l'un indépendant du locuteur (SI) et l'autre adapté au locuteur (SAT). Ces trois DNNs partagent la même topologie (exceptée la dimension des entrées) et sont entrainés avec les mêmes données. Un système GMM auxiliaire est également appris sur les mêmes données. L'apprentissage du modèle DNN SAT est décrit dans la Section 2 et la valeur de τ est fixée empiriquement à 5. L'apprentissage du modèle DNN SI avec les paramètres GMMD suit le processus décrit par la Figure 1 sans appliquer l'étape d'adaptation au locuteur. Les trois modèles disposent de 6 couches cachées à 2048 neurones et une couche de sortie de dimension 2355 correspondant aux états dépendant du contexte obtenus par une classification hiérarchique pour le système CD-GMM-HMM. Les DNN sont initialisés en empilant des machines de Boltzman restreintes. L'apprentissage final optimise une fonction d'entropie croisée et se termine par 5 itérations d'apprentissage séquentiel discriminatif avec un critère sMBR (*state Minimum Bayes Risk*).

Les expériences d'adaptation sont réalisées de façon non-supervisée sur les données d'évaluation en utilisant les transcriptions ou les graphes obtenus après la première passe. Deux expériences sont réalisées en adaptant le modèle GMM auxiliaire selon un critère : (1) *maximum a posteriori* standard, (2) *maximum a posteriori* utilisant les scores de graphes et la mesure de confiance décrite en Section 3 pour laquelle le seuil de confiance est fixé à $0,6$. Les performances des différentes adaptations sont reportées dans le Tableau 1 en terme de taux d'erreur mot pour les systèmes avec et sans adaptation.

Type of Features	Adaptation	WER, %	Δ WER, %
11×39MFCC	SI	7,51	référence
GMMD	SI	7,83	-
GMMD	SAT (MAP alignment)	7,09	5,6
GMMD	SAT (MAP lattice-based)	6,93	8,4

TABLE 1 – Taux d'erreur mots (%) évalué sur la tâche WSJ0 **si_et_20** et amélioration relative Δ WER par rapport à la référence.

5 Conclusions

Dans cet article, nous avons étudié des paramètres dérivés de GMM introduits récemment pour l'adaptation de modèles acoustiques DNN-HMM. Notre avons présenté une amélioration de l'approche précédemment proposée en appliquant le concept d'adaptation au locuteur au modèle DNN appris sur des paramètres dérivés de GMM et utilisant une mesure de confiance. Nous avons adapté un système GMM auxiliaire avec un critère de *maximum a posteriori* pour adapter le modèle DNN. Les résultats expérimentaux préliminaires obtenus sur le corpus WSJ0 démontrent que pour une adaptation non supervisée, la méthode proposée diminue relativement le taux d'erreur mots de $8,4\%$ par rapport à un système DNN-HMM indépendant du locuteur appris sur des paramètres MFCC standards. Il est important de noter que dans l'approche proposée, l'adaptation MAP du modèle GMM auxiliaire peut être remplacée par d'autres méthodes. Cette méthode fournit donc un cadre général pour transférer les algorithmes d'adaptation des modèles GMM-HMM pour les systèmes DNN.

Remerciements

Ce travail a ete partiellement finance par la commission europeenne a travers le projet EUMSSI, sous le numero de contrat 611 057, dans le cadre de lappel FP7-ICT-2013-10.

Références

ABDEL-HAMID O. & JIANG H. (2013). Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, p. 7942–7946 : IEEE.

ALBESANO D., GEMELLO R., LAFACE P., MANA F. & SCANZIO S. (2006). Adaptation of artificial neural networks avoiding catastrophic forgetting. In *Neural Networks, 2006. IJCNN'06. International Joint Conference on*, p. 1554–1561 : IEEE.

DODDIPATLA R., HASAN M. & HAIN T. (2014). Speaker dependent bottleneck layer training for speaker adaptation in automatic speech recognition. In *Fifteenth Annual Conference of the International Speech Communication Association*, p. 2199–2203.

DUPONT S. & CHEBOUB L. (2000). Fast speaker adaptation of artificial neural networks for automatic speech recognition. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, p. 1795–1798 : IEEE.

ELLIS D. P. & REYES-GOMEZ M. (2001). Investigations into tandem acoustic modeling for the aurora task. In *Eurospeech 2001 : Scandinavia : 7th European Conference on Speech Communication and Technology : September 3-7, 2001, Aalborg Congress and Culture Centre, Aalborg-Denmark : proceedings*, p. 189–192 : ISCA-Secretariat.

GALES M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer speech & language*, **12**(2), 75–98.

GAUVAIN J.-L. & LEE C.-H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *Speech and audio processing, IEEE transactions on*, **2**(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 *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, p. I–I : IEEE.

GOLLAN C. & BACCHIANI M. (2008). Confidence scores for acoustic model adaptation. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, p. 4289–4292 : IEEE.

HINTON G., DENG L., YU D., DAHL G. E., MOHAMED A.-R., JAITLY N., SENIOR A., VANHOUCKE V., NGUYEN P., SAINATH T. N. *et al.* (2012). Deep neural networks for acoustic modeling in speech recognition : The shared views of four research groups. *Signal Processing Magazine, IEEE*, **29**(6), 82–97.

KANAGAWA H., TACHIOKA Y., WATANABE S. & ISHII J. (2015). Feature-space structural maplr with regression tree-based multiple transformation matrices for DNN.

LEE L. & ROSE R. C. (1996). Speaker normalization using efficient frequency warping procedures. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, p. 353–356 : IEEE.

LEI X., LIN H. & HEIGOLD G. (2013). Deep neural networks with auxiliary gaussian mixture models for real-time speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, p. 7634–7638 : IEEE.

LI B. & SIM K. C. (2010). Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems. p. 526–529.

LI J., HUANG J.-T. & GONG Y. (2014). Factorized adaptation for deep neural network. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, p. 5537–5541 : IEEE.

LIAO H. (2013). Speaker adaptation of context dependent deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, p. 7947–7951 : IEEE.

LIU S. & SIM K. C. (2014). On combining DNN and GMM with unsupervised speaker adaptation for robust automatic speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, p. 195–199 : IEEE.

NETO J., ALMEIDA L., HOCHBERG M., MARTINS C., NUNES L., RENALS S. & ROBINSON T. (1995). Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system.

OCHIAI T., MATSUDA S., LU X., HORI C. & KATAGIRI S. (2014). Speaker adaptive training using deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, p. 6349–6353 : IEEE.

PAUL D. B. & BAKER J. M. (1992). The design for the wall street journal-based csr corpus. In *Proceedings of the workshop on Speech and Natural Language*, p. 357–362 : Association for Computational Linguistics.

PINTO J. P. & HERMANSKY H. (2008). *Combining evidence from a generative and a discriminative model in phoneme recognition*. Rapport interne, IDIAP.

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* : IEEE Signal Processing Society.

RATH S. P., POVEY D., VESELỲ K. & CERNOCKỲ J. (2013). Improved feature processing for deep neural networks. In *INTERSPEECH*, p. 109–113.

SAON G., SOLTAU H., NAHAMOO D. & PICHENY M. (2013). Speaker adaptation of neural network acoustic models using i-vectors. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, p. 55–59 : IEEE.

SEIDE F., LI G., CHEN X. & YU D. (2011). Feature engineering in context-dependent deep neural networks for conversational speech transcription. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, p. 24–29 : IEEE.

SENIOR A. & LOPEZ-MORENO I. (2014). Improving DNN speaker independence with i-vector inputs. In *Proc. ICASSP*, p. 225–229.

SINISCALCHI S. M., LI J. & LEE C.-H. (2013). Hermitian polynomial for speaker adaptation of connectionist speech recognition systems. *Audio, Speech, and Language Processing, IEEE Transactions on*, **21**(10), 2152–2161.

STADERMANN J. & RIGOLL G. (2005). Two-stage speaker adaptation of hybrid tied-posterior acoustic models. In *ICASSP (1)*, p. 977–980.

SWIETOJANSKI P. & RENALS S. (2014). Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, p. 171–176 : IEEE.

TOMASHENKO N. & KHOKHLOV Y. (2014). Speaker adaptation of context dependent deep neural networks based on MAP-adaptation and GMM-derived feature processing. In *Fifteenth Annual Conference of the International Speech Communication Association*, p. 2997–3001.

TOMASHENKO N. & KHOKHLOV Y. (2015). GMM-derived features for effective unsupervised adaptation of deep neural network acoustic models. In *Sixteenth Annual Conference of the International Speech Communication Association*, p. 2882–2886.

TRMAL J., ZELINKA J. & MÜLLER L. (2010). Adaptation of a feedforward artificial neural network using a linear transform. In *Text, Speech and Dialogue*, p. 423–430 : Springer.

UEBEL L. & WOODLAND P. C. (2001). Improvements in linear transform based speaker adaptation. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 1, p. 49–52 : IEEE.

XUE J., LI J., YU D., SELTZER M. & GONG Y. (2014). Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, p. 6359–6363 : IEEE.

YAO K., YU D., SEIDE F., SU H., DENG L. & GONG Y. (2012). Adaptation of context-dependent deep neural networks for automatic speech recognition. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, p. 366–369 : IEEE.

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 *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, p. 7893–7897 : IEEE.

Extraction automatique de contour de lèvre à partir du modèle CLNF

Li Liu[1,2] Gang Feng[1,2] Denis Beautemps[1, 2]

(1) Univ. Grenoble Alpes, GIPSA-lab, F-38040 Grenoble
(2) CNRS, GIPSA-lab, F-38040 Grenoble

li.liu@gipsa-lab.grenoble-inp.fr, gang.feng@gipsa-lab.grenoble-inp.fr,
denis.beautemps@gipsa-lab.grenoble-inp.fr

RESUME

Dans cet article nous proposons une nouvelle solution pour extraire le contour interne des lèvres d'un locuteur sans utiliser d'artifices. La méthode s'appuie sur un algorithme récent d'extraction du contour de visage développé en vision par ordinateur, CLNF pour *Constrained Local Neural Field*. Cet algorithme fournit en particulier 8 points caractéristiques délimitant le contour interne des lèvres. Appliqué directement à nos données audio-visuelles du locuteur, le CLNF donne de très bons résultats dans environ 70% des cas. Des erreurs subsistent cependant pour le reste des cas. Nous proposons des solutions pour estimer un contour raisonnable des lèvres à partir des points fournis par CLNF utilisant l'interpolation par spline permettant de corriger ses erreurs et d'extraire correctement les paramètres labiaux classiques. Les évaluations sur une base de données de 179 images confirment les performances de notre algorithme.

ABSTRACT

Automatic lip contour extraction using CLNF model.
In this paper a new approach to extract the inner contour of the lips of a speaker without using artifices is proposed. The method is based on a recent face contour extraction algorithm developed in computer vision. This algorithm, which is called Constrained Local Neural Field (CLNF), provides 8 characteristic points (landmarks) defining the inner contour of the lips. Applied directly to our audio-visual data of the speaker, CLNF gives very satisfactory results in about 70% of cases. However, errors exist for the remaining cases. We offer solutions for estimating a reasonable inner lip contour from the landmarks provided by CLNF based on spline to correct its bad behaviors and to extract the suitable labial parameters A, B and S. The evaluations on a 179 image database confirm performance of our algorithm.

MOTS-CLES : modèle CLNF, spline, contour des lèvres, paramètres labiaux, parole visuelle.
KEYWORDS: CLNF model, spline, lip contour, lip parameters, visual speech.

1 Introduction

Cet article traite de l'extraction du contour interne des lèvres à partir d'enregistrements vidéo du visage « naturel » (c.à.d. sans utilisation d'artifices) vu de face dans le contexte du traitement automatique de la parole. Ce contour constitue en effet une étape indispensable pour obtenir les

paramètres portant l'information visuelle de la parole en suivant une approche forme par opposition à une approche d'apparence (voir par exemple Potamianos et al., 2012). Les bénéfices de l'information visuelle pour la perception de la parole (lecture labiale) sont bien connus. Depuis les travaux de Sumby et Pollack (1954), à ceux de Benoit et collègues pour la langue française (1992) en passant par Summerfield et collègues (Summerfield, 1979 ; Summerfield et al., 1989), il est bien établi que l'information fournie par le mouvement du visage (principalement celui des lèvres), est utilisée pour améliorer la perception de la parole dans des situations de bruit ambiant. Les expériences en shadowing (répétition de la parole de l'autre) ont montré le bénéfice de l'apport de la collaboration audiovisuelle en situation de parole « audio claire » (Reisberg et al., 1987 ; Scarbel et al., 2014). L'effet McGurk manifeste dans le cas où les informations audio et vidéo sont incohérentes la capacité d'intégrer ces informations par l'identification d'un percept différent de celui porté par chacune des deux modalités seules (McGurk and MacDonald, 1976; MacDonald and McGurk, 1978). Le contour des lèvres et les paramètres labiaux (étirement A, aperture B et aire S) qui en sont extraits sont très utiles en traitement automatique et plus particulièrement en décodage visuo-phonétique par la reconnaissance de l'articulation labiale, domaine qui connait un regain d'intérêt pour les enjeux en surveillance ou en communication avec les sourds, réel enjeu de santé publique. Historiquement pour extraire ces contours, les lèvres étaient maquillées en bleu avant l'enregistrement vidéo. Le contour interne des lèvres était alors obtenu par application d'un simple seuil dans le plan « bleu » de l'image codée RGB (Lallouache, 1990 ; Lallouache, 1991 ; Aboutabit et al., 2007).

Plusieurs travaux ont eu pour objectif de s'affranchir de l'utilisation de maquillage des lèvres. Ainsi dans le domaine de la parole, Ming et al. (2010) ont proposé d'estimer directement les paramètres labiaux par les coefficients d'une décomposition en Cosinus Discrets à 2 dimensions de la région d'intérêt des lèvres non maquillées. Dans le domaine du traitement des images les approches s'appuyant sur des modèles de contour actif multi-paramétrés ont permis pour les plus récentes méthodes de segmenter les lèvres en ajustant les contours par plusieurs polynômes d'ordre trois et l'application de seuils multiples sur le paramètre de luminance pour ce qui concerne le contour interne (Stillitano et al. 2012). Dans le domaine de la vision par ordinateur, les méthodes s'appuient sur des modèles de formes et d'apparence et le modèle CLNF se situe dans ce contexte.

L'objectif de notre travail est d'appliquer cette dernière méthode issue du domaine de la vision par ordinateur au traitement automatique de la parole dans le domaine visuel (en phonétique, analyse/synthèse, reconnaissance audio-visuelle) afin d'extraire le contour interne des lèvres sans utilisation de quelque artifice expérimental que ce soit. Nous étudions ses performances en tant que méthode générique et proposons des améliorations pour tenir compte des spécificités rencontrées en parole. Nous montrons que cette approche permet d'obtenir de manière efficace les paramètres labiaux des lèvres sans passer par des modèles de lèvres complexes.

2 La base de données visuelles

Les données sont composées d'images vidéo vues de faces de voyelles extraites d'un corpus de 50 mots isolés du Français prononcés par un sujet et précédemment enregistré dans le contexte d'un travail sur la Langue Parlée Complétée (Ming et al., 2010). On obtient des images toutes les 20 ms. L'enregistrement audio synchrone a permis de segmenter les voyelles. Dans ces intervalles on sélectionne 2 à 3 images successives correspondant à la partie stationnaire de chaque voyelle. Enfin, afin d'équilibrer la base de données, nous avons fait un tri de ces voyelles dont la répartition est donnée dans la Table 1.

L'ensemble a permis de constituer une base de 179 images. Le contour interne des lèvres a été extrait manuellement par un expert, un des auteurs. L'expert a placé une quarantaine de points décrivant fidèlement le contour interne des lèvres tout en s'assurant qu'il s'agit du contour au sens

« articulatori-acoustique ». Les paramètres (A, B, S) sont extraits à partir du contour selon la méthode classique en parole visuelle (Lallouache, 1990, 1991) et exprimés en cm ou cm^2. En traçant dans le plan (A, B), nous avons bien observé la répartition en trois groupes classiques (groupe I : voyelles ouvertes et étirées aux lèvres, groupe II : voyelles ouvertes et arrondies, groupe III : voyelles fermées et arrondies).

	Groupe I					Groupe II			Groupe III				
Voyelle	[a]	[ɛ]	[ɛ̃]	[e]	[i]	[ã]	[ɔ]	[œ]	[o]	[ø]	[õ]	[y]	[u]
Effectif	26	18	15	21	21	24	12	12	6	6	6	3	9

TABLE 1 : Répartition des voyelles dans la base de données et leur répartition en trois groupes

3 Le modèle CLNF

En vision par ordinateur, l'algorithme AAM *Active Appearance Models* (Cootes et al., 1998) introduit un modèle statistique conjoint de forme (ensemble de points placés sur le visage) et d'apparence en niveaux de gris du visage vu de face ainsi qu'un algorithme d'ajustement linéaire du modèle sur les visages. Le modèle CLM *Constrained Local Model* (Cristinacce and Cootes, 2006) applique le modèle conjoint pour générer des *templates* (imagettes rectangulaires d'une dizaine de pixels centrées sur les 68 points du modèle) qui estiment les points à partir d'une relation non linéaire. En effet, les *templates* sont utilisés pour trouver les bords des segments du visage en optimisant une fonction de réponse de surface sous une contrainte de forme. Enfin le modèle CLNF *Constrained Local Neural Field* (Baltusaitis et al., 2013) est une amélioration du modèle CLM avec l'estimation des *templates* par la méthode LNF (*Local Neural Fields*) et l'utilisation d'une fonction d'optimisation s'appuyant sur la méthode *Non-Uniform RLMS*. Dans la méthode LNF, la probabilité d'une position étant donné le *template* est calculée à partir d'une distribution sigmoïde dont les paramètres sont estimés par un réseau de neurones artificiels à noyaux convolués. *Non-Uniform RLMS* (Saragih et al., 2011) est une méthode pour minimiser une fonction de coût composée du terme RLMS dont le terme en jacobien est pondéré par la matrice de covariance des *templates*. Enfin, le modèle CLNF a été construit à partir de 4000 visages vus de face extraits des bases de données indépendantes du locuteur HELEN, LFPW et Multi-PIE (communication personnelle de Tadas Baltrusaitis). Le CLNF améliore l'estimation des *templates* du module LNF et le module d'optimisation « non-uniform » RLMS.

4 Problèmes rencontrés et solutions proposées

Dans notre expérience, pour chaque voyelle retenue dans la base de données, la méthode CLNF a été appliquée aux images de chacun des mots contenant la voyelle considérée. Nous avons retenu les points correspondants aux lèvres en particulier les 8 points du contour interne pour chacune des images de la voyelle considérée. Lors de l'application de la méthode sur notre base de données, nous avons constaté qu'elle donne d'excellents résultats pour le contour interne des lèvres dans environ 70% des cas, et ce malgré un nombre relativement faible de points (8 points seulement). Nous montrons dans la figure 1.a un exemple.

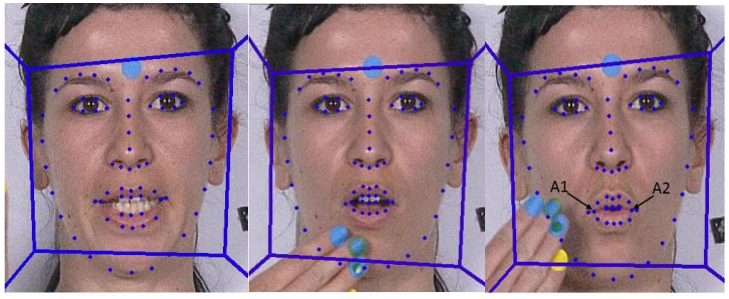


FIGURE 1: Illustration du résultat de l'application directe du modèle CLNF sur des visages de la base de donnée. On observe les 68 points distribués sur l'ensemble du visage et en particulier ceux de la région des lèvres. De gauche à droite, (a) montre des points bien placés, (b) les contours interne (et aussi externe) de la lèvre inférieure mal déterminés, (c) les points extrêmes du contour interne largement au-delà de l'ouverture des lèvres.

Cependant, une partie des images de la base de données présente des défauts manifestes. En particulier, on peut constater que le contour de la lèvre inférieure est parfois très mal déterminé (Figure 1.b). On n'a remarqué aucun problème pour la lèvre supérieure. Ce phénomène peut être expliqué par le fait que le modèle CLNF s'appuie sur un dictionnaire d'imagettes (*templates*). Si pendant la phase d'apprentissage, la région des lèvres n'a pas été bien prise en compte, il peut manquer des imagettes lors de la phase d'optimisation. Ce phénomène affecte davantage la lèvre inférieure car la région concernée est souvent très complexe (langue et des dents pouvant être partiellement visibles, voir Figure 1). Par ailleurs, nous avons constaté que pour les lèvres arrondies de faible ouverture, comme pour les voyelles [u], [y] par exemple, les deux points A1 et A2 marquant les extrémités horizontales du contour interne peuvent être bien éloignés du véritable contour au sens de celui de la relation articulatori-acoustique (Figure 1.c). En effet, d'un point de vue « géométrique », A1 et A2 ne sont pas faux car dans ce cas le contour interne peut réellement atteindre ces deux points d'extrémité. Cependant, d'un point de vue articulatori-acoustique, ces deux points ne définissent pas le paramètre d'étirement aux lèvres.

Pour avoir une idée claire sur les performances objectives de la méthode, nous comparons les points déterminés par CLNF avec le contour déterminé par l'expert. Cette comparaison se fait en termes des paramètres A, B et S, et non par une erreur quadratique moyenne entre deux contours. Le problème est délicat : comment estimer un contour à partir de seulement 8 points fournis par la méthode CLNF? On peut naturellement effectuer une interpolation linéaire permettant de relier deux points adjacents pour former un contour. Mais cette méthode présente des erreurs assez importantes vis-à-vis du véritable contour. Ceci étant dit, pour une première comparaison CLNF - contour d'expert, nous avons adopté l'interpolation linéaire car l'objectif premier est de déceler les erreurs importantes de la méthode CLNF et de les corriger. Nous proposons par la suite une interpolation non linéaire mieux adaptée au contour des lèvres. Une fois le contour déterminé, nous calculons les paramètres A, B et S. Nous traçons ensuite ces paramètres pour le contour estimé à partir des points CLNF et pour le contour d'expert, ainsi que leurs écarts (Figure 2). Nous constatons que les erreurs de CLNF concernant la lèvre inférieure mal déterminée sont clairement traduites par le paramètre B : leurs valeurs sont en général beaucoup plus petites que celles du véritable contour (Figure 2.b). En revanche, pour les erreurs correspondant aux lèvres arrondies de petite ouverture, c'est le paramètre A qui trahit l'anomalie (Figure 2.a).

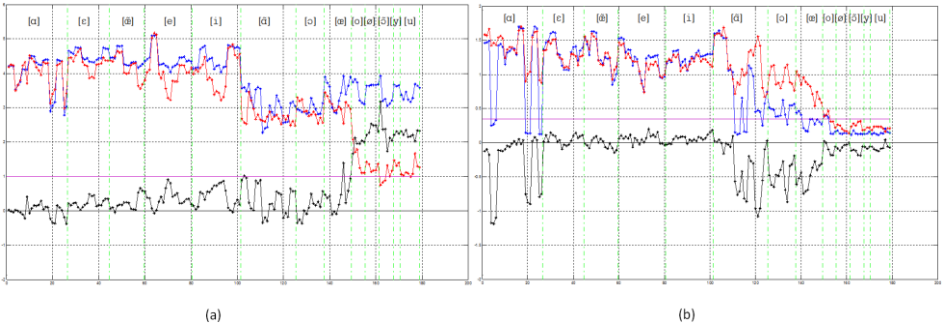


FIGURE 2 : Tracés de la valeur des paramètres A (a), B (b) issus du contour interne des lèvres déterminé par l'expert (tracé rouge) et produit par le modèle CLNF après interpolation linéaire (tracé bleu), et leur écart (tracé noir). Le trait de couleur mauve horizontal indique l'erreur quadratique moyenne. Les pointillés verticaux de couleur verte délimitent les voyelles concernées.

4.1 Solutions proposées

Etant donné la complexité de la méthode CLNF, en particulier sa phase d'apprentissage (4000 différents visages ont été utilisés), il n'est pas possible pour le moment de corriger les erreurs constatées en intervenant directement dans l'algorithme de base. Nous cherchons ainsi à corriger les défauts de l'algorithme par d'autres moyens. Ainsi nous proposons une estimation de la partie inférieure des lèvres basée sur le contraste de la luminance de l'image. En effet, sur la partie centrale des lèvres, lorsque l'on passe de la langue (ou des dents) à la lèvre inférieure, la luminance varie sensiblement. La frontière correspond à la variation la plus grande de la luminance dans la direction verticale. Certes, la recherche d'un champ de gradient dans la zone concernée permettrait de déceler les variations dans toutes les directions. Mais appliquée à des images assez bruitées cela ne donne pas de résultats convaincants. Ainsi nous décidons de chercher la position des extrema de la dérivée de la luminance dans la direction verticale, dans un intervalle délimité par les points fournis par le modèle pour la lèvre inférieure (Figure 3). Ce calcul de dérivée, très sensible au bruit, est envisageable seulement si on lisse préalablement les données. Ainsi nous proposons un lissage par spline avec un poids correctement choisi (p=0,1 valeur déterminée expérimentalement) (Feng, 1998). L'avantage de cette méthode est que la dérivée est obtenue naturellement lors du lissage. La figure 3 montre un exemple de ce lissage ainsi que la dérivée correspondante. Nous vérifions que le minimum de la dérivée correspond bien à la position de la lèvre inférieure.

Testée sur de nombreuses images contenant des erreurs de la méthode CLNF, l'estimation de la lèvre inférieure par détection de la plus grande variation de la luminance donne des résultats tout à fait satisfaisants. Il faut cependant savoir quand cette correction est nécessaire. Pour cela, nous constatons que le rapport A/B calculé à partir des points fournis par le modèle CLNF initial constitue un excellent indicateur. En effet, avec les erreurs de la CLNF, le paramètre B est anormalement petit, de telle sorte que A/B devient anormalement grand. Sachant que ce rapport A/B est relativement stable (typiquement entre 2 et 5 comme observé à travers tout le corpus), une valeur anormalement grande de A/B permet de déceler les erreurs. Ainsi nous effectuons la correction quand ce rapport est supérieur à 5. Les résultats d'évaluation par la suite confirment parfaitement cette valeur. A noter que l'utilisation d'un seuil ici n'apparait que comme indice de détection d'une anomalie de CLNF.

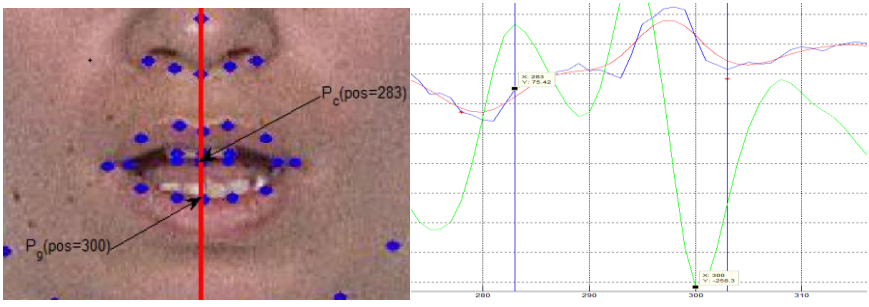


FIGURE 3 : Image illustrant une erreur de placement du contour interne de la lèvre inférieure (les points bleus prévus pour la partie inférieure se situent vers la lèvre supérieure) (à gauche). A droite, on présente la valeur de la luminance le long de la ligne verticale rouge (tracé bleu) et son lissage par spline (tracé rouge), ainsi que la dérivée du tracé lissé (tracé vert). On peut observer que le point du contour interne des lèvres correspond au minimum de la dérivée. Les 2 traits pleins verticaux de couleur bleue délimitent l'intervalle de recherche.

Nous avons évoqué que l'interpolation linéaire ne donne pas de résultats satisfaisants pour le contour estimé à partir des points CLNF, étant donné la très grande distance qui sépare les deux points d'extrémité (A1 et A2) et les autres points. Nous proposons l'utilisation d'une interpolation spline. Nous constatons que l'application de cette méthode, simple mais efficace, donne des contours excellents pour la partie inférieure des lèvres. En revanche, les trois points au centre de la lèvre supérieure forment souvent un « V », ce qui rend une interpolation spline aberrante dans le contour. Par ailleurs nous constatons que pour la partie supérieure, l'interpolation linéaire donne des résultats assez convenables, nous avons décidé de conserver l'interpolation linéaire pour la lèvre supérieure. La figure 4.a illustre le résultat d'un contour obtenu par cette méthode.

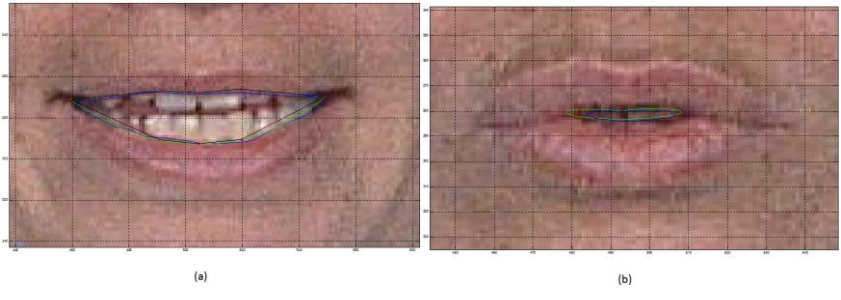


FIGURE 4 : A gauche, image illustrant l'effet de l'interpolation non linéaire. Le contour d'expert en tracé bleu, les 8 points de couleur rouge fournis par le modèle CLNF avec leur interpolation linéaire (tracé noire) et l'interpolation par spline concernant la partie inférieure (tracé vert). A droite le contour d'expert (tracé bleu), les 8 points de couleur rouge fournis par le modèle CLNF et l'interpolation par spline pour le contour interne entier mais excluant les 2 point extrêmes qui sont totalement erronés (tracé vert).

Pour les lèvres arrondies de petite ouverture, les deux points d'extrémité déterminés par CLNF sont manifestement faux. Une interpolation (linéaire ou spline) ne peut corriger cette anomalie. Remarquons que dans ce cas, les six points issus de CLNF (trois points supérieurs et trois points inférieurs) sont déterminés correctement. Nous proposons ainsi une estimation du contour uniquement

à partir de ces six points. Les deux points d'extrémité étant ignorés, on ne peut estimer un contour que si on considère la partie supérieure et la partie inférieure comme étant un ensemble et non deux parties séparées. Nous proposons une interpolation du contour entier à partir de ces six points. Voici la méthode proposée. Nous dilatons d'abord l'échelle verticale de telle sorte que les 4 coins de ces 6 points forment un carré. On convertit ensuite les coordonnées cartésiennes en coordonnées polaires pour le contour entier afin d'assurer la continuité du lissage aux deux extrémités. On effectue ensuite une interpolation avec ce système de coordonnées. Après l'interpolation, on revient à l'échelle initiale et on obtient un contour entier interpolé. Les résultats obtenus sont satisfaisants et leur évaluation est présentée dans la section suivante. Un exemple d'une telle interpolation est illustré à la figure 4.b. Ce traitement est uniquement appliqué aux lèvres arrondies de petite ouverture caractérisées par un très fort rapport A/B issu de la méthode CLNF supérieur à 8.

5 Evaluation des résultats et discussion

Nous avons évalué la méthode CLNF incluant nos propositions d'amélioration en utilisant la même base de données contenant 179 images. Nous montrons les paramètres A, B et S obtenus en intégrant toutes les propositions, comparées naturellement avec (A, B, S) du contour fourni par l'expert, ainsi que leur écart. Les résultats sont présentés à la figure 5 (ne concernant que les paramètres A et B). On constate que la correction des erreurs de la lèvre inférieure issues de CLNF donne des résultats totalement satisfaisants. En effet, dans la figure 5.b, on peut constater que toutes les erreurs présentes dans la figure 2.b ont été corrigées. Les valeurs de B suivent assez bien celles du contour de l'expert avec un écart cohérent avec les zones où on n'effectue pas cette correction. Rappelons que les résultats de la figure 5.b correspondent déjà à un contour de la lèvre inférieure interpolée par spline. Mais cette interpolation modifie très peu le paramètre B. Donc la différence entre la figure 5.b et la Figure 2.b résulte essentiellement de la correction sur la lèvre inférieure. On peut constater que l'erreur quadratique moyenne (toutes les voyelles confondues) passe de 0,3 cm pour la figure 1.b à 0,1 cm pour la figure 5.b, montrant l'efficacité de l'amélioration.

Examinons maintenant l'amélioration apportée par une interpolation du contour entier pour des lèvres arrondies de faible ouverture. Nous savons que ceci affecte essentiellement le paramètre A car CLNF donne les deux points d'extrémité beaucoup trop distants. On peut constater dans la figure 5.a que le paramètre A pour ces voyelles est beaucoup plus raisonnable, réduisant considérablement les écarts. Notons que le paramètre A n'est affecté ni par la correction sur la lèvre inférieure, ni par une interpolation non linéaire du contour inférieur, la différence entre la figure 5.a et la figure 2.a concerne uniquement les voyelles de ce groupe. L'erreur quadratique moyenne passe de 1,0 cm à 0,4 cm.

On peut remarquer que le paramètre A pour le groupe III reste supérieur à la valeur du contour expert. Ceci est essentiellement dû au fait que les six points issus de CLNF présentent en général une distance entre eux peu variable quelques soient les voyelles, ce qui constitue une limite de la méthode pour les très faibles ouvertures aux lèvres.

Nous examinons le paramètre S. La correction apportée au contour se répercute naturellement sur le paramètre S. La valeur de ce paramètre suit maintenant très bien celle issue du contour d'expert, et les écarts sont encore plus homogènes. L'erreur quadratique moyenne concernant S passe finalement de 0,89 cm^2 à 0,35 cm^2.

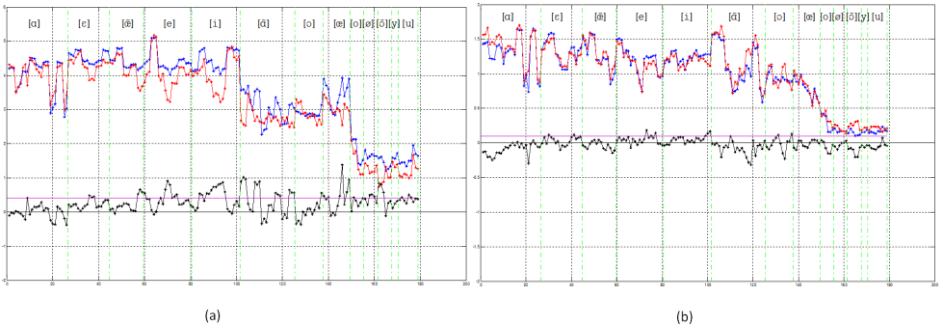


FIGURE 5 : Tracés de la valeur des paramètres A (a) et B (b) issus du contour interne des lèvres déterminé par l'expert (tracé rouge) et produit par le modèle CLNF incluant les 3 propositions de correction (tracé bleu), et leur écart (tracé noir). Le trait de couleur mauve horizontal indique l'erreur quadratique moyenne. Les pointillés verticaux de couleur verte délimitent les voyelles concernées.

Dans l'ensemble l'erreur absolue obtenue est homogène que l'on considère les grandes ou petites ouvertures aux lèvres, ce qui est caractéristique de la méthode. En conséquence l'erreur relative augmente considérablement pour les petites ouvertures ce qui diminue son intérêt. Mais pour pouvoir comparer avec la littérature nous avons reporté dans la table 2 les erreurs relatives qui se trouvent plus faibles que celles obtenues par Stillitano et al. (2013).

Erreur relative moyenne et (écart-type)	A : 13,4% (16%)	B : 9,8% (12%)	S : 13,6% (13%)

TABLE 2 : Erreur relative moyenne en % et écart-type pour les paramètres labiaux

6 Conclusion

En conclusion, nous retenons que le modèle CLNF issu du domaine de la vision par ordinateur et développé pour l'extraction de partie du visage complet reste très prometteur pour des données visuelles en production de parole. En effet, ce modèle générique permet d'extraire les paramètres du contour interne correctement dans environ 70% des cas sans aucune intervention spécifique. Nous avons montré pour le reste, qu'il a fallu développer des méthodes visant à corriger les erreurs en s'appuyant sur les points centraux fournis par le modèle CLNF. En effet, la correction a consisté à les repositionner sur le véritable contour en cherchant le maximum de contraste sur la luminance, en utilisant une interpolation spline ainsi que sa dérivée. Et dans le cas des petites ouvertures, nous avons pu proposer une interpolation spline de l'ensemble du contour interne s'appuyant sur les 6 points centraux issus du modèle CLNF. Les performances atteignent une précision de 1 mm pour le paramètre d'aperture B, de 4 mm pour le paramètre d'étirement A et de 0,35 cm^2 pour l'aire intéro-labiale S en terme d'erreur quadratique moyenne. Ces résultats sont comparables à ceux de la littérature mais sont obtenus à partir de lèvres sans maquillage. Ils indiquent que le modèle générique CLNF est tout à fait approprié. Enfin les améliorations apportées ici ne touchent pas le cœur du modèle CLNF et ses propriétés. Comme perspectives, il restera à élargir le corpus de données, en intégrant la variété des unités de parole et la variabilité liée à plusieurs locuteurs. On s'intéressera aussi à l'application de cette méthode à des situations plus complexes telles que par exemple les occlusions main-visage ou les variations dans les conditions d'enregistrement.

Références

ABOUTABIT N. (2007). Reconnaissance de la Langue Française Parlée Complétée. Manuscrit de thèse, Université de Grenoble.

AUER E.T., BERNSTEIN L.E. (2007). Enhanced Visual Speech Perception in Individuals With Early-Onset Hearing Impairment. *Journal of Speech, Language, and Hearing Research*, 50, 1157-1165.

Baltrusaitis T., Morency L.-P., and Robinson P. (2013). Constrained local neural fields for robust facial landmark detection in the wild. In *Computer Vision Workshops (ICCV-W)*, Sydney, Australia, 2013 IEEE Conference on. IEEE, 2013.

BENOIT C., LALLOUACHE T., MOHAMADI T., ABRY C. (1992). A set of French visemes for visual speech synthesis. In: Bailly G., Benoit C. (Eds.), Talking Machines: Theories, Models and Designs. Elsevier Science Publishers, Amsterdam, pp. 485-504.

COOTES TF , EDWARDS G.J., TAYLOR C.J. (1998). Active Appearance Model. *Actes de European Conference on Computer Vision*, 484-498.

CRISTINACCE D. AND COOTES T. (2006). Feature detection and tracking with Constrained Local Models. *Actes de British Machine Vision Conference, Vol. 3,* 929-938.

FENG G. (1998). Data Smoothing by Cubic Spline Filters. *IEEE Transactions on Signal Processing*, 46, 2790-2796.

LALLOUACHE T. (1990). Un poste Visage-Parole. Acquisition et traitement des contours labiaux. *Actes des Journées d'Etudes de la Parole*, Montréal.

LALLOUACHE T. (1991). Un poste Visage-Parole couleur. Acquisition et traitement automatique des contours des lèvres. Thèse de doctorat, Institut National Polytechnique de Grenoble.

MCGURK AND MACDONALD J, 1976. "Hearing lips and seeing voices", Nature 264, 746-748.

MACDONALD J., MCGURK H., 1978. Visual influences on speech perception processes. *Perception and Psychophysics* 24, 253-257.

MING Z., BEAUTEMPS D., FENG G. AND SCHMERBER S. (2010). Estimation of Speech Lip Features From Discrete Cosine Transform. Interspeech proceedings. Tokyo, Japan.

REISBERG D., MCLEAN J., GOLDFIELD A. (1987). Easy to hear but hard to understand: a lipreading advantage with intact auditory stimuli. In: Dodd, R., Campbell, R. (Eds.), Hearing by Eye : The Psychology of Lipreading. Lawrence Erlbaum Associates Ltd, Hillside, NJ, pp. 97-113.

POTAMIANOS, G., NETI, C., LUETTIN, J., AND MATTHEWS I. (2012). Audiovisual automatic speech recognition. In G. Bailly, P. Perrier, E. Vatikiotis-Bateson (Eds), *Audiovisual Speech Processing*, pp. 193-247.

Saragih, J., Lucey, S. and Cohn, J. Deformable Model fitting by Regularized Landmark Mean-Shift. *IJCV*, 2011.

SCARBEL L., BEAUTEMPS D., SCHWARTZ J.-L. & SATO M. (2014). The shadow of a doubt? Evidence for perceptuo-motor linkage during auditory and audiovisual close-shadowing. Front. Psychol.

STILLITANO S., GIRONDEL V., CAPLIER C. (2013). Lip contour segmentation and tracking compliant with lip-reading application constraints.

FN5, un modèle psycholinguistique informatique de la reconnaissance des mots parlés chez l'auditeur français, mis à la disposition des chercheurs et enseignants

Nicolas Lévy
Institut des Sciences du langage et de la communication,
Université de Neuchâtel, Pierre-à-Mazel 7, CH-2000 Neuchâtel, Suisse
nicolas.lewy@unine.ch

RÉSUMÉ

Voici un modèle psycholinguistique informatique pour le français. Il s'appelle FN5 et simule la reconnaissance humaine de mots parlés, présentés seuls (déterminant, adjectif antéposé, substantif) ou en suites de deux mots (déterminant et substantif, adjectif antéposé et substantif). Le modèle contient un lexique de 17 668 mots et cela dans deux versions, française et Suisse romande. Grâce à une architecture connexionniste localiste à trois niveaux (traits distinctifs, phonèmes, mots) qui est enrichie de plusieurs innovations clés (processeur de position, groupements de connexions, et point d'isolation), le modèle peut reconnaître la plupart des mots et des suites qu'on lui présente (taux de succès entre 83.6% et 99.7%), et en plus, il est capable de reproduire un grand nombre d'effets trouvés lors d'études expérimentales (ex. fréquence, longueur, effacement du schwa, liaison, etc.). Le modèle, qui possède une interface graphique, est téléchargeable, et utilisable à la fois pour la recherche et pour l'enseignement.

ABSTRACT

FN5, a computational psycholinguistic model of spoken word recognition in French, made available to researchers and teachers.

Here is a computational psycholinguistic model for French. It is called FN5 and simulates human recognition of spoken words, presented either alone (determiner, prenominal adjective, noun) or in two-word sequences (determiner and noun, prenominal adjective and noun). The model contains a lexicon of 17,668 words, in a standard French and a Swiss French version. Owing to a localist connectionist architecture of three levels (features, phonemes, words), enriched with several key innovations (position processor, groups of connections, and isolation point), the model is able to recognize most of the words and sequences presented to it (success rate between 83.6% and 99.7%), and it can also replicate a large number of effects found in experimental studies (e.g. frequency, length, schwa deletion, liaison, etc.). The model has a graphical interface, and can be downloaded and used both for research and teaching.

MOTS-CLÉS : reconnaissance humaine de mots parlés, simulation sur ordinateur, langue française, utilisation de modèles, interface graphique.

KEYWORDS: human spoken word recognition, computer simulation, French language, use of models, graphical user interface.

1 Introduction et motivation

La psycholinguistique informatique est une branche de recherche à l'intersection de la psychologie, la linguistique et l'informatique, et représente ainsi un exemple typique de pluridisciplinarité dans les sciences cognitives dont elle fait partie. Elle a pour but de simuler sur ordinateur, aussi clairement et honnêtement que possible, le traitement du langage et de la parole chez l'être humain (voir, par exemple, Crocker, 1996 ; Dijkstra et De Smedt, 1996, comme ouvrages de référence). Les avantages d'une modélisation informatique sont manifestes. Alors que la modélisation verbale est forcément imprécise, incomplète, et souvent ambiguë, la modélisation sur ordinateur oblige à rendre explicite tous les mécanismes du modèle. En outre, elle permet de visualiser sur l'écran le processus psycholinguistique en question. Mais en revanche elle demande, bien sûr, un important travail de programmation.

Jusqu'à récemment, les modèles proposés pour simuler le processus de la reconnaissance humaine de mots parlés, comme, entre autres, TRACE (McClelland et Elman, 1986), Shortlist A (Norris, 1994), Shortlist B (Norris et McQueen, 2008), et aussi ARTWORD (Grossberg et Myers, 2000), se sont concentrés sur l'anglais (ou, dans un cas, le néerlandais). Bien que les principes de base de la reconnaissance humaine de mots parlés restent, plus ou moins, les mêmes pour toutes les langues (aspects universaux), ce processus psycholinguistique doit aussi prendre en compte de nombreux et d'importants facteurs – tel que le contenu du lexique, les caractéristiques de ses mots, le répertoire des sous-unités, etc. – qui diffèrent considérablement d'une langue à une autre (aspects spécifiques aux langues ; voir Cutler, 2012). En conséquence, il est indispensable d'établir des modèles pour des langues autres que l'anglais.

FN5, le modèle psycholinguistique informatique que nous avons élaboré ces dernières années (Lévy *et al*., 2005 ; Lévy, 2015, 2016), est le premier modèle (et, à ce jour, le seul) à porter sur la reconnaissance des mots parlés chez l'auditeur français. Sur ces quelques pages, nous familiarisons le lecteur avec ce modèle : son lexique, son fonctionnement, son interface, son utilisation, etc. Maintenant que FN5 est complet, que son évaluation systématique est terminée avec succès, et que le modèle s'avère robuste, stable et fiable, il est prêt à être mis à la disposition de la communauté scientifique (voir les informations ci-dessous). Il nous semblerait donc que cette édition conjointe des JEP-TALN soit un moment idéal pour présenter notre modèle – et, bien évidemment, en faire une petite démonstration – à l'attention d'un public de spécialistes francophones.

2 Lexique du modèle

FN5 – le numéronyme n'a pas de signification – simule la reconnaissance de mots présentés seuls (déterminant, adjectif antéposé, substantif) ou en suites de deux mots (déterminant + substantif, adjectif antéposé + substantif). Son lexique est composé de 16 971 substantifs, 679 adjectifs et 18 déterminants, pour un total de 17 668 mots, et est fourni en deux versions, française et Suisse romande. (Dans la version Suisse romande, nous avons ajouté aux 35 phonèmes du français standard le /ɑ/, le /œ̃/, et 7 voyelles longues, /i: y: e: ɛ: ø: a: u:/, pour un total de 44 phonèmes.)

Pour constituer ce lexique, nous avons extrait l'ensemble des déterminants, adjectifs et substantifs de Brulex (Content *et al*., 1990) avant de les soumettre à toute une série de traitements à la fois informatiques et linguistiques. Nous avons éliminé les mots de certaines catégories (ex. substantifs au pluriel, adjectifs postposés uniquement, doublons, variantes orthographiques, mots de très basse fréquence, etc.), nous avons corrigé les indices de fréquence des mots à l'aide de Lexique 3 (New,

2006) et les avons normalisés (de sorte qu'ils s'étendent de 0 pour très rare à 1 pour très fréquent), et nous avons bien sûr fait vérifier l'ensemble de l'information linguistique présente dans le lexique (orthographe, transcription phonétique et genre). Nous avons également ajouté des informations sur les voyelles longues en finale de mot pour le français de Suisse romande (ex. « envie ») et sur les consonnes de liaison des adjectifs antéposés (ex. « grand » avec /t/, « heureux » avec /z/, etc.). Enfin, aux mots qui contiennent un schwa (ex. « fenêtre », « souvenir », etc.), qu'il soit obligatoire dans la prononciation, facultatif, ou interdit (2 190 mots en tout ; voir Racine et Grosjean, 2005), nous avons associé des indices de préférence de la variante avec et sans effacement du schwa.

3 Fonctionnement du modèle

Le modèle se sert du formalisme des réseaux connexionnistes localistes (Grainger et Jacobs, 1998). Dans la figure 1, nous présentons l'architecture générale du modèle qui s'inspire des modèles antérieurs – anglophones – de la reconnaissance de mots parlés, notamment de TRACE et Shortlist. Le modèle consiste en un grand nombre d'unités de base qui travaillent toutes en parallèle. Ces unités sont organisées en trois niveaux linguistiques – traits distinctifs, phonèmes, mots –, et sont connectées par des liens d'activation (visualisés par les flèches pointues) et/ou d'inhibition (visualisés par les flèches rondes). Il existe des liens ascendants, des traits distinctifs vers les phonèmes ainsi que des phonèmes vers les mots, et des liens descendants, des mots vers les phonèmes (optionnels dans le modèle). De plus, il y a des liens latéraux, aux niveaux des phonèmes et des mots. L'entrée du modèle consiste en une suite de valeurs au niveau des traits distinctifs, valeurs qui représentent le mot (ou la suite de mots) à reconnaître. C'est en passant par les divers liens d'activation et d'inhibition, et au moyen d'un certain nombre de cycles de simulation, que les traits distinctifs répercutent leur état d'activation vers les phonèmes concernés, et que ces phonèmes interagissent à leur tour avec les mots qui les contiennent.

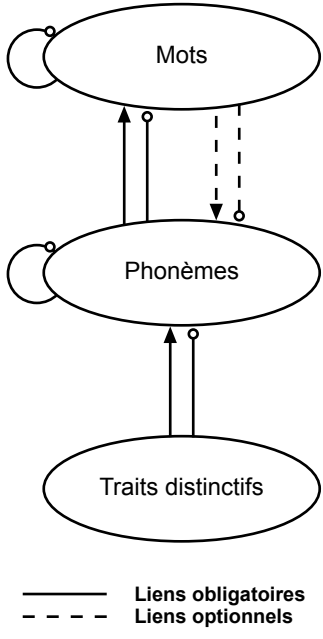


FIGURE 1 :
Architecture générale de FN5

Dans ce qui suit, nous présentons trois innovations clés.

3.1 Processeur de position

Afin de permettre la simulation de la reconnaissance d'une suite de deux mots (ex. « ta table » – /tatabl/), nous avons prévu qu'à chaque cycle, un processeur de position aligne les mots du lexique avec chaque position possible de la chaîne des phonèmes en entrée. Le modèle aligne donc le début du mot avec le début de la suite, à savoir, dans l'exemple /tatabl/, avec le /t/ initial en position 1, mais également avec le /a/ en position 2, avec le deuxième /t/ en position 3, etc. Il établit pour chacune des positions la somme combinée des activations et inhibitions du mot, compare ces sommes, et choisit la position qui produit la somme la plus élevée pour ce mot. C'est cette position qui est alors retenue et stockée avec le mot pour la suite de la simulation. Ce positionnement optimal du mot est répété à chaque cycle d'un phonème (il y a 16 cycles par phonème) et cela pour chaque phonème de la chaîne phonétique. Certes, la position du mot peut changer d'un cycle à l'autre, mais nous imposons au système de n'en avoir qu'une seule, pour chaque mot, à chaque

cycle. Si plusieurs positions donnent exactement le même résultat, le processeur de position choisit la position la plus récente. Bien entendu, dans un traitement séquentiel où les phonèmes arrivent les uns après les autres (d'abord le premier phonème, ensuite le deuxième phonème, etc.), le processeur de position ne peut placer un mot que dans la position 1 lors de l'arrivée du premier phonème, et aucune comparaison n'est effectuée à ce moment-là. C'est à l'arrivée du deuxième phonème qu'il peut placer le mot en position 1 ou en position 2 en comparant les deux sommes, et c'est à l'arrivée du troisième phonème qu'il compare les sommes des positions 1, 2 et 3, etc. De la sorte, l'activation et la reconnaissance d'un mot peuvent avoir lieu à n'importe quel emplacement de la chaîne parlée.

Lorsque nous comparons cette approche avec celle d'autres modèles qui simulent la reconnaissance de deux ou plusieurs mots (ex. TRACE et Shortlist), nous y voyons plusieurs avantages réunis en un seul modèle :

— L'approche fonctionne sans qu'il faille dupliquer les mots à toutes les positions théoriquement possibles : les mots n'ont qu'une seule position qui leur est attribuée à chaque cycle et ils peuvent être déplacés vers une autre position si le processeur de position le décide.
— L'approche ne limite pas le nombre des mots-candidats qui sont pris en considération ; tous les mots du lexique peuvent être activés en parallèle, et ce avec un niveau d'activation approprié (forte lors d'une bonne correspondance, moindre lors d'une correspondance partielle, etc.).
— L'approche ne nécessite pas que la chaîne en entrée soit segmentée préalablement ; elle s'appuie plutôt sur le fait que les frontières entre mots émergent souvent du processus normal de la reconnaissance des mots.

3.2 Groupements de connexions

Pour les mots qui contiennent des phonèmes qui peuvent être effacés (ex. un schwa) et/ou ajoutés (ex. une consonne de liaison), et plus généralement pour tous les mots avec plusieurs formes (comme les adjectifs), nous faisons appel aux groupements de connexions. Chacune des prononciations réalisables d'un mot (ex. avec ou sans schwa, avec ou sans consonne de liaison, masculin ou féminin, etc.) possède un groupement de connexions qui relient le mot avec les phonèmes en question. Par exemple, pour le substantif « pelouse », il y a un groupement de connexions pour la forme avec effacement du schwa (/pluz/) et un autre pour la forme sans effacement (/pəluz/). De même, pour l'adjectif « grand », il existe un groupement pour la forme féminine (/gʁɑ̃d/), un deuxième pour la forme masculine sans liaison (/gʁɑ̃/), et un troisième pour la forme masculine avec liaison (/gʁɑ̃t/).

Le processeur de position n'aligne donc pas seulement chaque mot avec chaque position possible dans la chaîne des phonèmes en entrée, mais aussi, dans le cas de plusieurs groupements de connexions, il établit les sommes des activations et inhibitions pour chaque groupement, à chaque position, afin de trouver le groupement et la position qui donnent la somme la plus élevée (ex. le groupement qui correspond à la forme masculine avec liaison, et cela pour la position 2). À la fin de la comparaison, les informations concernant le groupement et la position sont stockées avec le mot. Le processus recommence à chaque nouveau cycle et peut donc produire un résultat différent, le processeur de position ne retenant qu'un seul groupement et qu'une seule position pour chaque mot. Notons que des informations numériques peuvent être attachées aux groupements de connexions afin de rendre compte de la préférence, ou de la fréquence, d'un groupement donné chez l'être humain (voir les mots avec schwa facultatif, par exemple, où il y a souvent une préférence pour l'une ou l'autre version).

3.3 Point d'isolation

Le point d'isolation (ou PI) est le moment, exprimé en nombre de cycles, où le mot à reconnaître dépasse le niveau d'activation de tous les autres candidats possibles ; de plus, ce premier rang doit être maintenu jusqu'à la fin de la simulation. Cette mesure est calculée pour chaque mot de la suite (dans le cas d'une suite de mots), et elle est proposée en nombre de cycles ainsi qu'en pourcentage de la longueur du mot (neutralisant ainsi la longueur). Le PI est une mesure simple, mais fort utile dans la simulation de la reconnaissance de mots parlés.

4 Un exemple de simulation

Dans la figure 2, nous montrons la fenêtre principale de l'interface utilisateur graphique du modèle qui comporte deux parties. Celle du haut, avec divers boutons, zones de texte, etc., sert à régler les différentes composantes du modèle et à entrer les mots pour lesquels on veut simuler la reconnaissance ; celle du bas, libellée « Activation/Cycle », sert à visualiser l'activation des mots.

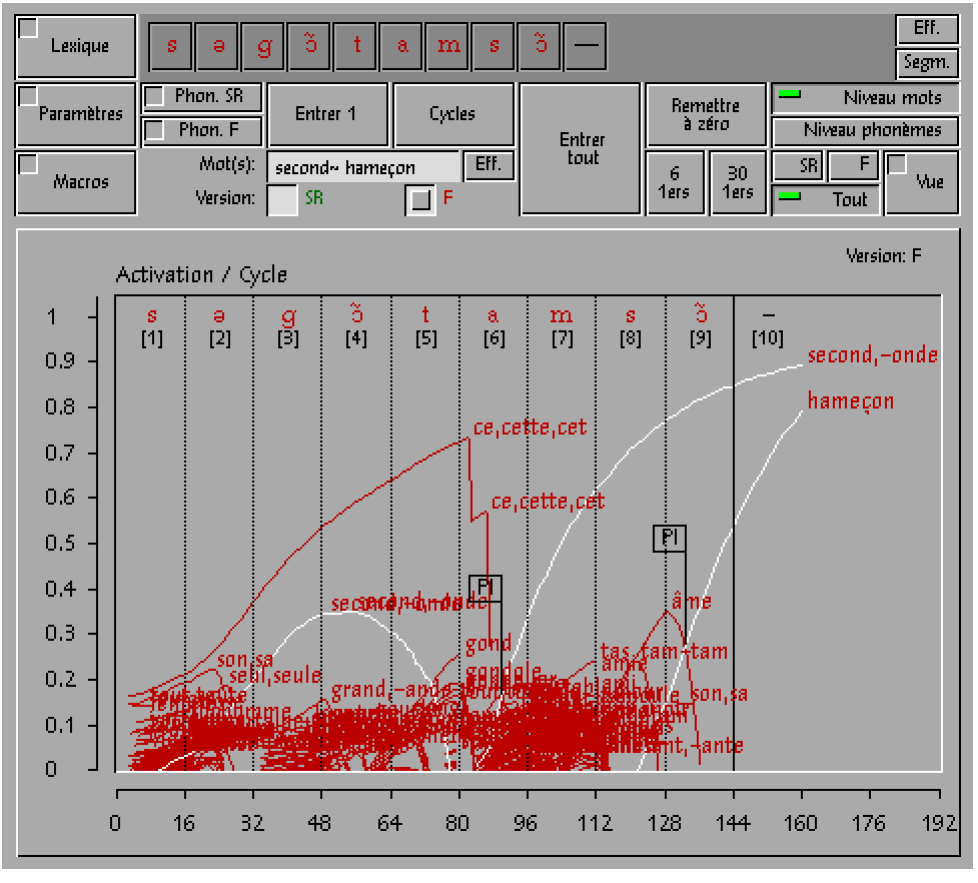


FIGURE 2 : Fenêtre de l'interface de FN5 avec la simulation décrite dans le texte

À titre d'exemple, nous montrons la simulation de la reconnaissance de la suite « second hameçon » (/səgɔ̃tamsɔ̃/), prononcée avec un schwa dans « second » mais sans schwa dans « hameçon » et, en outre, avec une liaison entre les deux mots. Pour obtenir le résultat présenté dans la figure, nous avons d'abord chargé le lexique du modèle dans sa version française (voir le bouton « Lexique », en haut à gauche, qui permet de choisir le fichier « FrenchPtitami.Lex »). Ensuite, nous avons tapé les mots « second hameçon » dans la zone de texte intitulée « Mot(s) ». La transcription phonétique /səgɔ̃tamsɔ̃–/ est alors fournie par le logiciel dans la zone en haut, à droite du bouton « Lexique », comme on le voit dans la figure. Notons que le /–/ marque la fin de la chaîne. Quant à la liaison, elle se manifeste visuellement de deux manières différentes : au niveau de la suite phonétique, un /t/ est inséré automatiquement au bon endroit, et dans la suite graphique (zone « Mot(s) »), un signe « ~ » apparaît entre « second » et « hameçon » lorsque l'on fait un retour à la ligne après avoir tapé les mots. Quant à la présence ou absence du schwa, la transcription proposée est celle avec la plus haute préférence dans le lexique, mais cela reste à notre choix d'y insérer ou effacer un schwa. Enfin, nous avons cliqué sur « Entrer tout », le bouton qui sert à insérer les phonèmes dans le modèle, un par un, en commençant avec le premier phonème et en finissant par la marque de fin de suite. L'état d'activation des mots est alors apparu dans la zone « Activation/Cycle » de la figure, et a évolué, comme on l'observe dans la figure, au fur et à mesure de l'avancement dans la suite.

Ce qui rend cet exemple intéressant est qu'au début du processus (sur la gauche donc de la fenêtre), le modèle s'engage sur une fausse piste, mais qu'il réussit à s'en tirer (vers le centre de la fenêtre), et qu'il nous propose les deux mots justes à la fin du processus (sur la droite de la fenêtre). En effet, à la fin des cinq premiers phonèmes de la suite, /s/, /ə/, /g/, /ɔ̃/ et /t/ (au cycle 80 donc), le modèle a activé les mots « ce, cette, cet » et « gond » (ainsi que « gondole », « gondolier », etc.) au détriment de « second, -onde » qui est tombé dans le négatif (voir la courbe blanche qui représente le niveau d'activation du mot entré). C'est pendant l'arrivée du prochain phonème, du /a/ (cycles 81-96), que le modèle se rend compte qu'il se trouve dans une impasse (il n'y a aucun mot dans son lexique qui commence par /gɔ̃ta/). Le modèle désactive donc « ce, cette, cet » en faveur de « second, -onde » qui est capable de récupérer et de bientôt dépasser les autres candidats ; en fait, le point d'isolation (PI) de « second, -onde » se trouve au cycle 90 (voir le petit drapeau « PI »). Alors sont également activés les mots qui débutent par /ta/ (ex. « table », « tableau », « tas », etc., avec le /t/ rattaché au substantif) ou par /a/ (ex. « ami », « amour », « âme », « affaire », etc., quand le /t/ est une consonne de liaison). Encore pendant l'entrée du /m/ (cycles 97-112), il y a de nombreux candidats possibles, d'où la masse de mots actifs. (Notons qu'en cliquant sur le bouton « 6 1^{ers} » (six premiers) ou « 30 1^{ers} » (trente premiers), on peut obtenir une liste des candidats les plus actifs, et ce à tout moment de la simulation.) Parmi ces derniers, le mot « hameçon » est de basse fréquence, et donc il sort assez tard dans le processus. Mais, puisque ce mot est le seul à correspondre aux derniers phonèmes de la suite, /s/ et /ɔ̃/ (cycles 113-144), ce mot émerge, après tout, comme le vrai gagnant (avec un PI au cycle 133 ; voir le deuxième petit drapeau).

5 Évaluation systématique

On vient de constater que le modèle est donc capable de reconnaître une suite de deux mots comme « second hameçon », ce qui est, bien évidemment, très positif et gratifiant. Mais, comme on vient de le remarquer également, en outre du (simple) succès de reconnaissance, le modèle nous permet de découvrir bien d'autres choses encore. À l'aide du modèle, nous étudions les mots-candidats qui jouent un rôle important pendant le processus de reconnaissance d'un mot seul ou d'une suite de mots ; nous observons que certains mots, ou certaines suites de mots, sont reconnus plus vite, et d'autres plus lentement ; nous mesurons le temps que le modèle prend pour reconnaître ces mots ou

ces suites (à savoir le nombre de cycles de simulation, en utilisant la mesure du PI) ; et nous cherchons à décrire les raisons pour lesquelles le modèle éprouve de la facilité ou de la difficulté. Même les échecs de reconnaissance du modèle, qu'ils soient momentanés ou complets (ces derniers pas trop nombreux, espérons-le), peuvent donc être utiles.

Dans ce qui suit, nous résumons les principaux éléments et résultats de notre toute dernière évaluation systématique du modèle (voir Lévy, 2015, pour les détails et les analyses statistiques).

Dans une première partie de cette évaluation, sur la reconnaissance générale du modèle, nous avons utilisé 1 000 mots uniques (que nous avons tirés au hasard de notre lexique) ainsi que 1 000 suites de deux mots (générées par un programme pseudo-aléatoire qui a toujours combiné un déterminant ou adjectif antéposé avec un substantif, en observant l'accord du genre et en faisant la liaison où elle s'applique, mais en ignorant la sémantique). Chaque mot après l'autre, et chaque suite de mots après l'autre, ont été testés dans le modèle (heureusement, nous avons accès à des fonctions de macro). Voici combien en sont reconnus par le modèle :

— 997 des mots uniques sont reconnus si le modèle peut savoir qu'il s'agit d'un mot unique, et
— 913 des mots uniques sont reconnus quand le modèle ne doit pas le savoir ;
— 993 des suites sont reconnues si le modèle peut savoir qu'il s'agit d'une suite de deux mots, et
— 836 des suites sont reconnues lorsqu'il ne doit pas le savoir.

Dans la deuxième partie de l'évaluation, composée de plusieurs études paramétriques de simulation avec FN5, nous nous sommes assurés que le modèle peut rendre compte de certaines variables propres au mot, ou propres à une chaîne de mots, et de leurs effets sur la reconnaissance humaine. Nous avons choisi des stimuli qui correspondent à deux niveaux d'une variable à tester (les autres variables étant contrôlées) ou nous avons adapté, lorsque cela était possible, des stimuli tirés d'une étude psycholinguistique expérimentale. En les testant dans le modèle, nous avons établi qu'il est capable de simuler les six effets suivants :

— les mots courts sont reconnus plus rapidement que les mots longs (effet de la longueur) ;
— les mots de haute fréquence sont reconnus plus promptement que les mots de basse fréquence (effet de la fréquence d'occurrence) ;
— les mots à point d'unicité précoce sont reconnus plus vite que les mots à point d'unicité tardif (effet du point d'unicité) ;
— les mots avec voyelle courte en finale sont reconnus plus tôt que les mots avec voyelle longue en finale (effet de la durée de la voyelle dans le français de Suisse romande) ;
— lorsque des mots sont prononcés avec schwa ils sont reconnus plus rapidement que lorsqu'ils sont prononcés sans schwa, si pour ces mots l'effacement du schwa est facultatif ou interdit mais non pas quand l'effacement est obligatoire (effet de l'effacement du schwa) ;
— des mots qui se trouvent dans une suite enchaînée avec liaison sont reconnus plus lentement que lorsqu'ils se trouvent tout seuls, mais ils sont reconnus tout aussi vite quand ils se trouvent dans une suite enchaînée sans liaison (effet de l'enchaînement avec liaison).

6 Disponibilité et utilisation du modèle

Le modèle FN5 est mis à disposition à des fins de recherche et d'enseignement académiques. Il est offert en combinaison avec BIMOLA (Lévy et Grosjean, 2008), un modèle psycholinguistique informatique de la reconnaissance des mots parlés chez le bilingue, mais on peut se servir de l'un ou de l'autre modèle, indépendamment.

Notre logiciel tourne sous OS X. Il est téléchargeable sur le site http://www.bimola.ch et s'installe vraiment très facilement et rapidement (contactez-nous si ce n'est pas le cas). La seule petite complication est qu'il fonctionne en tant qu'application XQuartz, et donc il vous faut tout d'abord installer XQuartz, si ce dernier n'est pas déjà présent sur votre Mac ; voir http://www.xquartz.org.

Notons que l'interface utilisateur du logiciel peut être basculée entre le français et l'anglais, en appuyant sur un seul bouton. Par ailleurs, en suivant les quelques instructions qui sont données dans la fenêtre de démarrage du logiciel (lancer le modèle, charger un lexique, taper le mot ou les mots à reconnaître, etc.), vous mettrez bientôt en route vos toutes premières simulations avec FN5. Pourquoi ne pas commencer par l'exemple de « second hameçon » décrit ci-dessus ?

6.1 Utilisation par le chercheur

Étant donné que FN5 est le premier et seul modèle psycholinguistique informatique à décrire la reconnaissance des mots parlés chez l'auditeur français, nous espérons qu'il se révélera comme un bon nouvel instrument pour les recherches menées en psycholinguistique sur le français. On voudra peut-être :

— s'appuyer sur le modèle simplement d'une manière théorique et l'utiliser pour expliquer le processus général de la reconnaissance des mots parlés ;
— se servir du modèle pour interpréter les résultats d'une étude expérimentale spécifique ;
— chercher à reproduire les résultats d'une étude expérimentale en la simulant dans le modèle ;
— explorer comment les stimuli d'une étude se comportent avant même de mener l'expérience ;
— valider le comportement du modèle avec des études expérimentales nouvelles et ciblées ;
— tester le fonctionnement du modèle en modifiant/branchant/bloquant certaines composantes du modèle et en examinant les conséquences ;
— générer de nouvelles prédictions à l'aide du modèle qui pourront ensuite être examinées dans des études expérimentales auxquelles on n'a pas encore pensé jusqu'ici.

6.2 Utilisation par l'enseignant

Grâce à sa vitesse et son interface utilisateur graphique, le modèle se prête bien à la démonstration en direct. Il peut donc servir de didacticiel dans des cours et séminaires en sciences du langage, psycholinguistique, phonétique, orthophonie, linguistique informatique, psychologie cognitive, etc.

Il est clair qu'un modèle fonctionnant, présenté en cours d'exécution et projeté en grand sur un tableau blanc, est un outil pédagogique très parlant. C'est sans doute un grand avantage pour l'usage éducatif que ce modèle concerne le français, et que l'on peut visualiser l'activation de phonèmes et de mots dans notre langue. Le lexique du modèle (avec ses 17 668 mots) se compare très favorablement au lexique du fameux modèle TRACE (qui n'avait que quelques centaines de mots). On se sert, selon ses besoins, de la version française ou Suisse romande. On peut soi-même ajouter à ce lexique tout mot éventuellement manquant que l'on désirerait faire reconnaître. Dans Léwy (2015), on trouve diverses boîtes pratiques permettant de se familiariser avec le fonctionnement détaillé du modèle, ainsi que des exemples dont on peut s'inspirer pour ses propres simulations.

Une dernière suggestion : à l'aide du bouton « Segm. » (en haut à droite dans l'interface), on peut se faire trouver toutes les manières possibles de segmenter une chaîne de phonèmes en suites de mots (tout en laissant de côté la syntaxe et sémantique). On y voit combien est difficile la tâche de reconnaissance de mots parlés !

Références

CONTENT A., MOUSTY P., RADEAU M. (1990). Brulex : une base de données lexicales informatisée pour le français écrit et parlé. *L'Année psychologique* 90, 551-566.

CROCKER M. (1996). *Computational psycholinguistics: An interdisciplinary approach to the study of language*. Dordrecht, Pays-Bas : Kluwer Academic.

CUTLER A. (2012). *Native listening: Language experience and the recognition of spoken words*. Cambridge, MA : MIT Press.

DIJKSTRA T., DE SMEDT K. (eds). (1996). *Computational psycholinguistics: AI and connectionist models of human language processing*. Londres : Taylor & Francis.

GRAINGER J., JACOBS A. (eds.). (1998). *Localist connectionist approaches to human cognition*. Mahwah, NJ : Erlbaum.

GROSSBERG S., MYERS C. (2000). The resonant dynamics of speech perception: Interword integration and duration-dependent backward effects. *Psychological Review* 107, 735-767.

LÉWY N. (2015). *Computational psycholinguistics and spoken word recognition in the bilingual and the monolingual*. Thèse de doctorat, Université de Neuchâtel. https://doc.rero.ch/record/257161

LÉWY N. (2016). *Simulating the recognition of isolated and connected words in spoken French: The FN5 model*. Manuscrit en préparation.

LÉWY N., GROSJEAN F. (2008). The Léwy and Grosjean BIMOLA model. In F. Grosjean, *Studying bilinguals*, p. 201-210. Oxford, Angleterre : Oxford University Press.

LÉWY N., GROSJEAN F., GROSJEAN L., RACINE I., YERSIN C. (2005). Un modèle psycholinguistique informatique de la reconnaissance des mots dans la chaîne parlée du français. *Journal of French Language Studies* 15, 25-48.

MCCLELLAND J., ELMAN J. (1986). The TRACE model of speech perception. *Cognitive Psychology* 18, 1-86.

NEW B. (2006). Lexique 3 : une nouvelle base de données lexicales. *Actes de la 13e Conférence sur le TALN (Traitement Automatique des Langues Naturelles)*, 892-900. Louvain, Belgique : ATALA.

NORRIS D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition* 52, 189-234.

NORRIS D., MCQUEEN J. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review* 115, 357-395.

RACINE I., GROSJEAN F. (2005). Le coût de l'effacement du schwa lors de la reconnaissance des mots en français. *Canadian Journal of Experimental Psychology* 59, 240-254.

Fusion d'espaces de représentations multimodaux pour la reconnaissance du rôle du locuteur dans des documents télévisuels

Sebastien Delecraz, Frederic Bechet, Benoit Favre, Mickael Rouvier
Laboratoire d'Informatique Fondamentale de Marseille
UMR 7279 CNRS / Aix-Marseille Université
163 avenue de Luminy, 13288 Marseille Cedex 9, FRANCE
{firstname.lastname}@lif.univ-mrs.fr

RÉSUMÉ

L'identification du rôle d'un locuteur dans des émissions de télévision est un problème de classification de personne selon une liste de rôles comme présentateur, journaliste, invité, etc. À cause de la non-synchronie entre les modalités, ainsi que par le manque de corpus de vidéos annotées dans toutes les modalités, seulement une des modalités est souvent utilisée. Nous présentons dans cet article une fusion multimodale des espaces de représentations de l'audio, du texte et de l'image pour la reconnaissance du rôle du locuteur pour des données asynchrones. Les espaces de représentations monomodaux sont entraînés sur des corpus de données exogènes puis ajustés en utilisant des réseaux de neurones profonds sur un corpus d'émissions françaises pour notre tâche de classification. Les expériences réalisées sur le corpus de données REPERE ont mis en évidence les gains d'une fusion au niveau des espaces de représentations par rapport aux méthodes de fusion tardive standard.

ABSTRACT

Multimodal embedding fusion for robust speaker role recognition in video broadcast

Person role recognition in video broadcasts consists in classifying people into roles such as anchor, journalist, guest, etc. Existing approaches mostly consider one modality, either audio (speaker role recognition) or image (shot role recognition), firstly because of the non-synchrony between both modalities, and secondly because of the lack of a video corpus annotated in both modalities. Deep Neural Networks (DNN) approaches offer the ability to learn simultaneously feature representations (embeddings) and classification functions. This paper presents a multimodal fusion of audio, text and image embeddings spaces for speaker role recognition in asynchronous data. Monomodal embeddings are trained on exogenous data and fine-tuned using a DNN on 70 hours of French Broadcasts corpus for the target task. Experiments on the REPERE corpus show the benefit of the embeddings level fusion compared to the monomodal embeddings systems and to the standard late fusion method.

MOTS-CLÉS : Identification du rôle du locuteur, fusion multimodale, émissions de télévision.

KEYWORDS: Speaker role recognition, multimodal speaker embeddings, broadcast news.

1 Introduction

L'identification du rôle d'une personne dans des émissions de télévisions consiste en un problème de classification de personne (parlant et/ou visible) selon une liste de rôles comme présentateur principal, journaliste, invité, etc. Dans ce contexte, les modalités audio et image sont naturellement complémentaires puisque l'on retrouve les caractéristiques du rôle dans le signal audio, la transcription écrite de la parole et les caractéristiques des scènes. De nombreuses approches proposées jusqu'à présent pour l'identification du rôle de la personne ne prennent en compte qu'une seule des modalités et ce pour deux principales raisons : premièrement la présence d'une personne n'est pas toujours synchronisée sur les différentes modalités. En effet, les locuteurs ne sont pas toujours visibles et tous les visages visibles à l'écran ne sont pas tous en train de parler. De plus, le manque de données multimodales annotées limite les possibilités pour réaliser un apprentissage joint de systèmes multimodaux qui supposent généralement une parfaite synchronie entre les modalités.

Les approches récentes basées sur les réseaux de neurones profonds (*Deep Neural Networks*, DNN) ont atteint des performances état-de-l'art pour de nombreuses tâches du traitement de l'audio et de l'image. Le principal avantage de ces techniques est d'apprendre simultanément des caractéristiques de représentations et des fonctions de classification. L'initialisation des caractéristiques de représentations peut être effectuée sur de grands corpus de données génériques pas nécessairement liés à la tâche cible pour plonger les données dans des espaces de représentations (dénommés *embeddings* en anglais) qui pourront être ajustés à la tâche cible de façon jointe. Cette approche a été proposée pour la tâche synchrone de détection et identification d'activité labiale (Ngiam *et al.*, 2011).

Dans cet article, nous voulons classifier les locuteurs en quatre rôles en utilisant les modalités audio, image et texte :

- R1 : les présentateurs, caractérisés par leur présence tout au long de l'émission ;
- R2 : les journalistes, des professionnels du monde de la télévision qui apparaissent une fois ou plus au cours d'une émission ;
- R3 : les reporters, proches du rôle R2, ce sont les correspondants qui couvrent les événements en dehors du plateau de l'émission ;
- R4 : les invités et autres, invités pour parler de l'actualité en raison de leur expertise ou leur renommée, ne prenant pas part à l'organisation et n'étant pas les leaders des débats ; les autres sont toutes les personnes qui peuvent apparaître, comme les personnes interviewées lors d'un reportage.

Nous présentons dans cet article [1] une alternative au paradigme de fusion tardive standard basée sur des *embeddings* multimodaux ajustés pour la tâche d'identification du rôle du locuteur (*Speaker Role Recognition*, SRR). La principale nouveauté de notre approche est une fusion au niveau des *embeddings* qui caractérise une information multimodale sans qu'une synchronie entre les modalités ne soit requise. Les expériences sur le corpus français REPERE met en évidence le gain de cette approche au regard de stratégies monomodales et des méthodes de fusion tardive.

2 Travaux connexes

L'identification automatique du rôle du locuteur (SRR) admet que les rôles soient identifiables par des caractéristiques spécifiques acoustiques, visuelles et textuelles comme le style de langage ou la

1. Partiellement traduit à partir de l'article (Rouvier *et al.*, 2015b).

prosodie. Dans la littérature, les méthodes de SRR ont été étudiées dans le but de catégoriser des documents audio-visuels (émissions de débats et d'informations). Les méthodes existantes sont séparées suivant les caractéristiques extraites (audio et/ou texte) ; le niveau de décision : pour chaque tour de parole (Liu, 2006) ou sur l'ensemble des tours de parole pour un locuteur donné (Dufour *et al.*, 2011; Hutchinson *et al.*, 2010; Zhang *et al.*, 2010; Wang *et al.*, 2011) ; et des techniques de classification supervisée (Dufour *et al.*, 2011; Bigot *et al.*, 2010; Liu, 2006) ou non-supervisée (Hutchinson *et al.*, 2010; Zhang *et al.*, 2010).

Dans (Dufour *et al.*, 2011), basé sur l'hypothèse que la classification de discours spontanés est un indice pour la tâche de SRR, les auteurs proposent une application de la détection de discours spontanés pour la tâche de SRR. Dans (Hutchinson *et al.*, 2010) les auteurs ont proposé un système non-supervisé de regroupement de locuteurs suivant leur rôle basé sur des caractéristiques structurelles et lexicales. Dans (Bigot *et al.*, 2010), les auteurs utilisent des caractéristiques temporelles, acoustiques et prosodiques pour classifier les rôles au niveau d'un groupe de locuteurs.

(Liu, 2006) classifie les rôles des locuteurs en utilisant des modèles de Markov cachés (*Hidden Markov Model*, HMM) et des classifieurs à maximum d'entropie (MaxEnt) avec une reconnaissance automatique de la parole (*Automatic Speech Recognition*, ASR) et une segmentation des tours de paroles des locuteurs réalisée manuellement. (Damnati & Charlet, 2011) ont présenté un système multimodal basé sur des caractéristiques lexicales et acoustiques pour la reconnaissance du rôle.

Pour ce qui est de la modalité visuelle, il n'y a pas à notre connaissance de travaux reposant sur l'image pour la détection du rôle des locuteurs. Les recherches portent surtout sur la reconnaissance de type de plan, comme par exemple dans (Feng *et al.*, 2014).

3 Approche

L'approche que nous proposons consiste en la création de représentation pour chaque modalité adaptées à la tâche de SRR. Ces représentation sont utilisées en entrée d'un classifieur multimodal qui peut tirer avantage de caractéristiques cross-modales tirées de la concaténation des représentations monomodales. La Figure 1 illustre cette approche.

Chaque représentation monomodale est entraînée sur un grand corpus monomodal qui n'est pas nécessairement en lié à la tâche de SRR. Le corpus multimodal annoté est utilisé seulement lorsque l'on entraîne la fusion. Cette méthode nous permet de tirer parti en même temps des fusions précoces et tardives : nous pouvons utiliser un grand corpus de données monomodales pour lequel nous n'avons pas une synchronie d'annotation dans les autres modalités comme cela peut être fait en fusion tardive ; nous entraînons des classifieurs multimodaux qui permettent la construction de caractéristiques multimodales directement depuis chaque modalité comme la fusion précoce.

Pour la modalité texte nous entraînons un réseau de neurones convolutionnel (*Convolutional Neural Network*, CNN) qui commence par l'apprentissage d'*embeddings* de mot sur un grand corpus de textes. La modalité audio utilise une représentation extraite d'un DNN modélisé à partir d'un système d'identification du locuteur, mais entraîné sur la tâche de SRR. La modalité image repose sur une représentation extraite d'un réseau de neurones d'identification de concepts visuels *ImageNet*, et raffiné pour la tâche cible. La fusion consiste en la concaténation des couches cachées de ces systèmes monomodaux à laquelle sont ajoutées des couches de neurones complètement reliées (*fully-connected*) pour créer un espace de représentation multimodal dans lequel sont fusionnées les caractéristiques

monomodales nécessaires à la prise de décision. La section suivante détaille l'architecture des composants monomodaux et multimodaux.

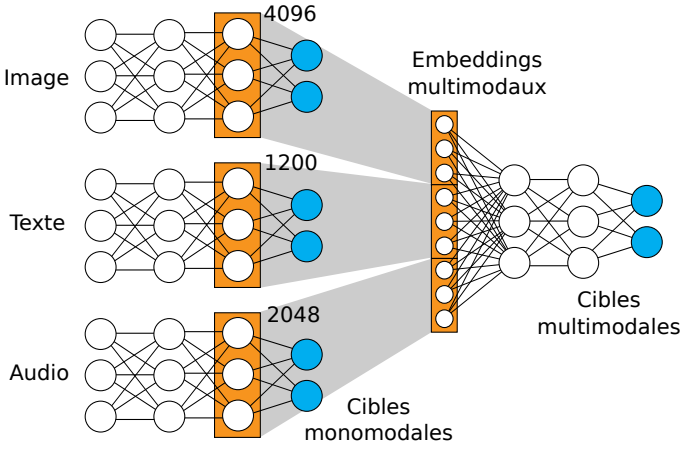


FIGURE 1 – Illustration de l'approche de fusion d'*embeddings*. Les systèmes monomodaux sont d'abord entraînés de manière indépendante, puis les activations de leurs couches cachées sont concaténées pour servir d'entrée au réseau multimodal. La taille des *embeddings* est donnée pour chaque modalité.

4 Modalité texte

De récents travaux ont montré que les CNN sont très performants dans le Traitement Automatique du Langage Naturel (TALN) pour des problèmes de classification (Collobert, 2011). Un CNN est un réseau de neurones profonds possédant plusieurs couches de convolution et de *max-pooling* suivis par un simple classifieur (souvent un Perceptron multicouche). Le principal avantage d'utiliser la convolution est la capacité du modèle de traiter des entrées de dimension variable (des phrases dans notre cas). De plus, les multiples filtres convolutionnels extraient des N-grammes lexicaux de différentes granularité alors que les couches de *pooling* extraient des caractéristiques sémantiques globale de l'entrée. Dans nos travaux, nous utilisons la transcription du discours du locuteur courant pour la tâche de SRR.

Premièrement, chaque mot est représenté par un vecteur de dimension 300 à valeurs réelles appelé *word embeddings*. Dans nos expériences, les *word embeddings* [2] ont été entraînés sur Wikipedia en utilisant le modèle *skip-gram* (5 itérations avec une fenêtre de taille 7). Cette stratégie nous permet de caractériser des associations grammaticales et sémantiques entre les mots.

Puis, les *word embeddings* pour les mots du tour de parole courant sont passé au travers de trois filtres convolutionnels qui sélectionnent les meilleurs 3-grammes, 4-grammes et 5-grammes. Ils sont combinés avec un une couche de *max-over-time pooling* (dimension 400) puis une couche de *soft-max fully-connected*. Nous utilisons du *dropout* pour désactiver 40% des neurones à chaque itération, agissant comme une régularisation.

2. Nous avons utilisé les toolkit *Word2Vec*.

5 Modalité audio

Dans nos travaux précédents, il a été proposé d'apprendre des caractéristiques acoustiques de haut niveau pour l'identification du locuteur (Rouvier *et al.*, 2015a), appelées *speaker embeddings*. Dans la même idée, nous proposons d'apprendre des caractéristiques du rôle du locuteur de haut niveau, appelées "*audio embeddings*", en utilisant des modèles profonds entraînés pour la tâche de SRR.

Les *audio embeddings* sont appris de la façon suivante : premièrement, un vecteur de caractéristiques acoustique de dimension 60 est extrait pour chaque tour de parole [3] avec un taux d'échantillonnage de 10ms (19 MFCCs, log énergie et deltas du premier et second ordre). Puis les statistiques du premier ordre, centrées-normalisées, obtenues depuis un modèle du monde (*Universal Background Model*, UBM) sont générées. Le modèle du monde est un GMM diagonal à 1024 composantes indépendant du canal (calculé avec le toolkit *Kaldi* [4]). Ensuite, les statistiques du premier ordre sont utilisées comme entrées d'un DNN avec deux couches cachées de 2048 neurones. Les fonctions d'activation de ces couches sont des unités linéaires rectifiées (ReLU) et la sortie est un *soft-max*. L'entraînement est effectué en minimisant l'entropie croisée sur les données d'apprentissage. Dans nos expériences, tous les hyper-paramètres des DNN sont déterminés sur un corpus de développement. La taille de mini-batch est de 512, 8 époques ont été effectuées et le taux d'apprentissage de 0.04 est réduit à 0.004 lors de la convergence. *In fine*, les *embeddings* audio de taille 2048 sont extraits à partir de la dernière couche cachée du DNN et utilisés pour la fusion multimodale.

6 Modalité visuelle

Les grammaires visuelles et les émissions de débats et d'informations sont une vraie source d'informations pour l'identification du rôle du locuteur.Nous utilisons des *embeddings d'image* issue de DNN qui se sont montrés extrêmement performants lors des campagnes d'évaluation *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC). Nous affinons le CNN nommé *AlexNet* (Krizhevsky *et al.*, 2012) entraîné sur le corpus de données ILSVRC-2012 pour la classification d'image. L'architecture du réseaux se compose de cinq couches de convolution, trois couches *fully-connected*, des couches de *max-pooling* et de normalisation. Elles utilisent la fonction *ReLU* (*Rectified Linear Unit*) comme fonction d'activation pour accélérer l'apprentissage et du *dropout* après les deux premières couches de neurones *fully-connected* pour éviter le sur-apprentissage. Ce modèle prend en entrée des images redimensionnées en (256×256) et normalisées en mini-batches de taille 512. Les poids sont mis à jour suivant les règles décrites dans (Krizhevsky *et al.*, 2012).

Nous avons adapté l'architecture *AlexNet* en changeant la dernière couche de neurones *fully-connected* pour prédire seulement quatre classes et ajusté les poids déjà appris du modèle *AlexNet* pour obtenir un nouveau CNN pour la tâche de SRR. Nous augmentons le taux d'apprentissage sur les couches de neurones *fully-connected* (dix fois le taux d'apprentissage global) afin de régulariser l'affinage de l'entraînement. Nous avons entraîné le réseau pendant 270 époques sur 19k images en utilisant *Caffe* [5] sur des GPU. Pour terminer, les *embeddings d'images* sont extraits depuis la seconde couche de neurones *fully-connected*, ce qui nous fournit des vecteurs de dimension 4096 qui seront utilisés dans le système de fusion multimodale.

3. La durée d'un segment moyen est de 7,8 seconde dans le corpus d'apprentissage.
4. http ://kaldi-asr.org/
5. http ://caffe.berkeleyvision.org/

7 Fusion multimodale

Il existe deux approches utilisées communément pour la fusion multimodale : les fusions précoce et tardive. La fusion tardive considère que les modalités sont indépendantes en appliquant d'abord des classifieurs séparés pour chaque modalité puis en fusionnant leurs sorties dans un classifieur de haut niveau. Malheureusement, le classifieur ne peut pas modéliser les corrélations entre les modalités et a seulement accès aux décisions des systèmes monomodaux. L'approche par fusion précoce contourne ce problème en apprenant des caractéristiques et des relations entre les classes pour modéliser les interactions entre les modalités. Toutefois, cette approche nécessite une synchronie entre les modalités.

Nous proposons une approche par fusion précoce basée sur des DNN où les entrées sont les *embeddings* de toutes les modalités pour la tâche spécifiée. Premièrement, les DNN sont entraînés indépendamment pour chaque modalité pour pouvoir en extraire des représentations monomodales générales (*embeddings* de texte, d'audio et d'image). Puis, ces *embeddings* sont utilisés en entrées d'un nouveau DNN entraîné pour apprendre à classifier le rôle des locuteurs à l'aide de caractéristiques multimodales. Contrairement à la fusion tardive, notre méthode peut tirer avantage de sous-espaces de caractéristiques pertinents (*embeddings*) de toutes les modalités.

Dans nos expériences, le DNN utilisé pour la fusion précoce est composé de deux couches cachées de dimension 1024. La non-linéarité de ces couches cachées est corrigée par une fonction d'activation ReLU. Les poids sont mis à jour en utilisant des mini-batch de taille 512, entraînés pendant 6 époques. Le taux d'apprentissage initialement de 0.01 est réduit à 0.001 lors de la convergence.

8 Expériences

Nous présentons les expériences réalisées sur le corpus multimodal REPERE (Giraudel *et al.*, 2012). La segmentation en locuteurs est effectuée par le système du LIUM (Rouvier *et al.*, 2013) qui obtient un Diarization Error Rate (DER) de 12.03% sur ce corpus. La transcription automatique est générée par *Kaldi* [4] et obtient un taux d'erreur mot (WER) de 19.67%. Le système est décrit en détail dans (Rouvier & Favre, 2014). Dans les émissions TV, les locuteurs apparaissent à l'écran seulement 60% du temps et les têtes visibles ne parlent que 30% du temps. Cet asynchronisme nous pousse à choisir une seule image pour représenter chaque tour de locuteur. Pour un tour de parole donné, il s'agit de l'image médiane de la plus grande intersection avec la segmentation en plan.

Les expériences sont menées sur le corpus REPERE (Giraudel *et al.*, 2012) rassemblant 70 heures d'émissions TV de 9 chaînes françaises. Chaque tour de parole est manuellement annoté avec la transcription, l'identité des locuteurs, et les rôles des locuteurs selon les classes suivantes : présentateur (R1), journaliste / chroniqueur (R2), reporter terrain (R3) et invité / autre (R4 [6]). Le corpus est divisé en ensembles d'apprentissage (18951 tours de parole), de développement (1402 tours de parole) et de test (4627 tour de parole) utilisés respectivement pour l'apprentissage des systèmes, la validation de la structure des réseaux de neurones et des hyper-paramètres des classifieurs, et l'évaluation des résultats. L'ensemble de test contient des types d'émissions qui se trouvent aussi dans les ensembles d'apprentissage et de développement (c'est le jour des émissions qui diffère), aussi bien que des nouveaux types d'émission qui sont inconnus des modèles entraînés sur l'ensemble d'apprentissage et

6. Fusion des rôles R4 et R5 du corpus original car cette paire de classes a un accord inter-annotateur faible.

de développement, permettant de vérifier la capacité de généralisation de notre modèle. La répartition des rôles dans l'ensemble de test et la suivante : 23,34% de présentateurs, 11,28% de journalistes, 14,22% de reporters et 51,16% d'autres.

Tous les résultats sont donnés en utilisant la précision (le nombre de rôles correctement identifié) et le *Diarization Error Rate* (DER). Le DER consiste à calculer les erreurs de SRR au niveau des trames, la même métrique étant utilisée pour les tâches de segmentation de locuteur. Le principal avantage de cette métrique est qu'elle nous permet de comparer deux sorties de systèmes de SRR différentes avec une segmentation en locuteur différente. Dans les résultats, DER-Man correspond au DER calculé sur la segmentation et transcription manuelles, alors que DER-Auto correspond à la segmentation et transcription automatiques.

Premièrement, la Table 1 compare les approches de bases et celles d'apprentissage profond monomodales. Parmi les approches de bases nous avons : *Majorité* où l'on choisit le rôle le plus fréquent pour chaque trame ; *Adaboost*, un classifier à base de *boosting* [7] sur des n-grammes de mots (modalité texte) ; *JFA* qui entraîne des modèles JFA sur les MFCC (*Joint-Factor-Analysis*) (Kenny *et al.*, 2005) (modalité audio) ; et *SVM-HOG* un classifieur à base de SVM sur des histogrammes de gradient (modalité image). Les résultats de la Table 1 indiquent clairement que les approches qui utilisent des DNN surpassent les systèmes de base. De plus, la modalité audio montre les classifieurs monomodaux les plus performants.

Système	**Modalité**	**Acc-Man**	**DER-Man**	**DER-Auto**
JFA	Audio	26,76	37,48	42,54
DNN-Audio	Audio	**77,52**	**19,79**	**25,43**
Adaboost	Texte	62,13	28,80	34,33
CNN-Texte	Texte	67,50	29,11	32,66
SVM-HOG	Image	62,76	36,97	42,04
CNN-Image	Image	70,48	25,69	35,25

TABLE 1 – Scores de précision et DER monomodaux sur l'ensemble de test.

Dans la suite, nos expériences montrent les résultats d'une fusion tardive basée sur un classifieur SVM. Toutes les probabilités données pour chaque modalité sont regroupées dans un vecteur, et un SVM linéaire est entraîné sur ces vecteurs de probabilités pour prédire le rôle du locuteur. Les résultats présentés dans la Table 2 montrent qu'une fusion des décisions au niveau des *embeddings* est plus performante qu'une fusion tardive. Ils justifient aussi l'utilisation de modèles multimodaux pour la tâche : le gain des performances des systèmes multimodaux comparé aux systèmes monomodaux est très important. Le meilleur score de DER-Man était de 19,79 (respectivement 25,43 pour le DER-Auto) pour les systèmes monomodaux alors qu'il est seulement de 13,84 (respectivement 19,9) pour les systèmes multimodaux. Nous pouvons aussi observer que c'est le modèle qui utilise les trois modalités qui donne les meilleurs résultats.

Afin d'étudier la robustesse de nos méthodes, la Table3 montre la précision et le DER sur un sous-ensemble du corpus de test qui correspond à des conditions d'émissions inconnues (le format des émissions est différent). Le système basé sur des *embeddings* de texte est robuste dans ces conditions alors que les résultats des modalités audio et image diminuent sur ces nouvelles émissions. C'est particulièrement vrai pour la modalité image qui passe d'un DER de 25,69 sur tout le corpus de test à 43,29 sur les émissions inconnues seulement. Dans ces conditions il n'est pas surprenant que les

7. https ://code.google.com/archive/p/icsiboost/

Système	Modalité	Acc-Man	DER-Man	DER-Auto
Majorité	-	51,16	39.77	44.54
Tardive	A+T	78,49	18,67	24,11
Tardive	A+I	80,98	17,26	22,98
Tardive	I+T	78,02	21,16	27,60
Tardive	A+I+T	82,36	15,37	20,97
Embedding	A+T	80,16	15,90	21,82
Embedding	A+I	82,16	15,45	2065
Embedding	I+T	76,01	22,83	28.60
Embedding	A+I+T	**85,28**	**13.84**	**19.79**

TABLE 2 – Scores de précision et DER pour les fusions tardives à posteriori et au niveau des *embeddings* (**A**udio, **I**mage, **T**exte).

méthodes de fusion ne donnent pas de meilleurs résultats que la meilleure des modalités seule. Ces résultats pointent une des faiblesses de notre approche quand toutes les modalités n'ont pas la même capacité de généralisation sur des événements inconnus.

Système	Modalité	Acc-Man	DER-Man	DER-Auto
CNN-Texte	T	**70,07**	**26,65**	**28,42**
DNN-Audio	A	65,69	29,88	37,31
CNN-Image	I	51,09	43,29	46,47
Tardive	A+I+T	70,07	27,77	34,61
Embedding	A+I+T	66,42	32,80	34,06

TABLE 3 – Résultats dans les conditions d'émission inconnue (5% de l'ensemble de test).

9 Conclusion

Dans cet article, nous avons introduit un système d'identification du rôle du locuteur basé sur la fusion d'espaces de représentations multimodaux pour des données asynchrones. Les expériences sur le corpus REPERE en utilisant une segmentation en locuteur et une transcription manuelles ou automatiques ont montré que la fusion de caractéristiques textuelles, audio et visuelles améliore considérablement les performances pour la classification en rôles des locuteurs au regard d'approches monomodales. Nos *embeddings* multimodaux ont permis de capturer les caractéristiques du rôle du locuteur sous plusieurs points de vues et l'utilisation d'une fusion au niveau des *embeddings* permet d'obtenir les meilleurs résultats avec 19,79% de DER sur une segmentation en locuteur automatique. Notre méthode tire avantage à la fois des fusions tardive et précoce en même temps : nous pouvons utiliser une grande quantité de données monomodales pour lesquelles nous n'avons pas d'annotations synchrones dans les autres modalités comme il se fait en fusion tardive ; nous entraînons des classifieurs multimodaux pour construire des caractéristiques multimodales directement depuis chaque modalité comme en fusion précoce.

Cependant, un des inconvénient de cette méthode est son manque de généralisation des modèles audio et image dans des conditions de contenus inconnus. L'augmentation de la robustesse dans ces conditions ainsi que l'application de notre modèle à d'autres tâches comme l'identification du locuteur sont des pistes intéressantes suite aux résultats obtenus.

Remerciements

Ces travaux ont été réalisés grâce à l'appui du projet A*MIDEX (n° ANR-11-IDEX-0001-02) financé par le programme du Gouvernement français "Investissement d'Avenir" et dirigé par l'Agence Nationale de la Recherche (ANR) ; ainsi qu'au soutien financier apporté par la Direction Générale de l'Armement (DGA) en partenariat avec Aix-Marseille Université dans le cadre du *Club des partenaires Défense*.

Références

BIGOT B., FERRANÉ I., PINQUIER J. & ANDRÉ-OBRECHT R. (2010). Speaker role recognition to help spontaneous conversational speech detection. In *SCSS*.

COLLOBERT R. (2011). Deep learning for efficient discriminative parsing. In *AISTATS*.

DAMNATI G. & CHARLET D. (2011). Multi-view approach for speaker turn role labeling in TV broadcast news shows. In *InterSpeech*.

DUFOUR R., ESTEVE Y. & DELÉGLISE P. (2011). Investigation of spontaneous speech characterization applied to speaker role recognition. In *Interspeech*.

FENG B., BAI J., CHEN Z., HUANG X. & XU B. (2014). Anchor shot detection with deep neural network. In *PCM*.

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

HUTCHINSON B., ZHANG B. & OSTENDORF M. (2010). Unsupervised broadcast conversation speaker role labeling. In *ICASSP*.

KENNY P., BOULIANNE G., OUELLET P. & DUMOUCHEL P. (2005). Factor analysis simplified. In *ICASSP*.

KRIZHEVSKY A., SUTSKEVER I. & HINTON G. E. (2012). Imagenet classification with deep convolutional neural network. In *NIPS*.

LIU Y. (2006). Initial study on automatic identification of speaker role in broadcast news speech. In *NAACL*.

NGIAM J., KHOSLA A., KIM M., NAM J., LEE H. & NG A. Y. (2011). Multimodal deep learning. In *ICML*.

ROUVIER M., BOUSQUET P.-M. & FAVRE B. (2015a). Speaker diarization through speaker embeddings. In *EUSIPCO*.

ROUVIER M., DELECRAZ S., FAVRE B., BENDRIS M. & BECHET F. (2015b). Multimodal embedding fusion for robust speaker role recognition in video broadcast. In *ASRU*.

ROUVIER M., DUPUY G., GAY P., KHOURY E., MERLIN T. & MEIGNIER S. (2013). An open-source state-of-the-art toolbox for broadcast news diarization. In *InterSpeech*.

ROUVIER M. & FAVRE B. (2014). Speaker adaptation of dnn-based asr with i-vectors : Does it actually adapt models to speakers ? In *InterSpeech*.

WANG W., YAMAN S., PRECODA K. & RICHEY C. (2011). Automatic identification of speaker role and agreement/disagreement in broadcast conversation. In *ICASSP*.

ZHANG B., HUTCHINSON B., WU W. & OSTENDORF M. (2010). Extracting Phrase Patterns with Minimum Redundancy for Unsupervised Speaker Role Classification. In *NAACL*.

L'impact des variations temporelles intrinsèques et extrinsèques de la voyelle sur la relation consonne-voyelle : Étude translinguistique sur l'arabe jordanien et le français

Mohammad Abuoudeh Olivier Crouzet
Laboratoire de Linguistique de Nantes UMR 6310 CNRS / Université de Nantes
Chemin de la Censive du Tertre 44312 NANTES
mohammad.abuoudeh@univ-nantes.fr, olivier.crouzet@univ-nantes.fr

RÉSUMÉ

Cette étude permet d'explorer les variations spectrales engendrées par deux types de variations temporelles qui résultent respectivement de l'opposition de longueur vocalique et des variations de débit de parole. Deux protocoles expérimentaux ont été conçus, l'un en arabe jordanien et l'autre en français, pour examiner ce phénomène. Un intérêt particulier a été porté aux occlusives produites dans des séquences CVC dans le but d'étudier la consonne en position initiale et la coarticulation anticipatoire. La durée des voyelles et la fréquence des trois premiers formants au début et au milieu de chaque séquence ont été mesurées dans chaque condition de longueur / débit. Les équations de locus ont été utilisées afin de décrire la relation CV quand elle subit ces deux types de variations. Selon les résultats, la qualité de la voyelle et de la consonne est influencée dans l'opposition de durée et dans le débit de parole. Ce changement généré par les variations temporelles est détecté à l'aide des équations de locus. Ces dernières révèlent qu'il existe un chevauchement coarticulatoire plus important quand la durée de la voyelle décroît.

ABSTRACT

The impact of extrinsic and intrisic vowel temporal variations on the consonant-vowel relationship : A trans-linguistic investigation on Jordanian arabic and French

This aims at examining the different spectral variations that are produced by two types of vowel time variations : phonological vowel length opposition and differing speaking rates. Two experiments on Jordanian Arabic and on French were conducted to investigate this phenomenon. We were interested in stop consonants produced in the initial position in CVC sequences. Vowel duration and the frequency of the first three formants were measured for each vowel length / speech rate. In addition, locus equations were computed to measure the impact of time variations on CV coarticulation.According to our results, it seems that both vowel length opposition and speaking rates had an impact on consonant and voyelle quality as well as on CV coarticulation. For the two time variation types, when the time decreases (more coarticulation overlap) locus equation slopes tend to have higher values.

MOTS-CLÉS : longueur vocalique, débit de parole, transitions formantiques, arabe jordanien, français.

KEYWORDS: vowel length, speaking rate, formant transitions, Jordanian Arabic, French.

1 Introduction

Les variations temporelles d'une voyelle sont associées à l'opposition de durée phonologique dans certaines langues. Cette opposition (voyelle longue vs. courte) est intrinsèquement liée à la voyelle, autrement dit les voyelles longues sont physiquement plus longues que les voyelles courtes. Les variations temporelles d'une voyelle peuvent aussi dépendre des variations de débit de parole. Ce phénomène extrinsèque à la voyelle change également la durée ; les voyelles produites à débit lent sont physiquement plus longues que les mêmes voyelles produites à débit rapide. L'objectif de ce travail est d'étudier l'impact de ces différents types de variations temporelles sur les qualités spectrales de la voyelle et de la consonne ainsi que sur la relation coarticulatoire entre la consonne et la voyelle. Plusieurs études ont été menées pour comprendre les différences entre voyelles longue et courte au niveau quantitatif (durée) et qualitatif (fréquence). Hadding-Koch & Abramson (1964) révèlent que le critère de quantité des voyelles n'est pas suffisant pour percevoir la distinction entre voyelles longue et courte en suédois. Dans la plupart des cas, la qualité des voyelles est indispensable pour différencier les deux. Dans une étude translinguistique sur l'arabe, le japonais et le thaï, Tsukada (2009) observe que l'influence de la quantité de la voyelle sur sa qualité est faible dans toutes les langues étudiées.

Il existe par ailleurs un débat à propos de l'impact du débit de parole sur les propriétés spectrales. Lindblöm (1963) observe que la qualité des voyelles (en suédois) est modifiée quand elles subissent des changements de débit. Il propose ainsi la notion d'*undershoot* : les voyelles n'atteignent pas leur cible articulatoire / acoustique quand elles sont confrontées à certains phénomènes dont la variation du débit de parole. Dans l'étude de Gay (1978), l'impact des variations de débit sur les propriétés spectrales des voyelles de l'anglais semble négligeable. Ce sont plutôt les consonnes (les F_{onsets}) qui seraient influencées par ce phénomène. Sur le même sujet d'étude, O'Shaughnessy (1986) a étendu le travail de Gay (1978) au français québecois en mettant en évidence que les variations temporelles attachées au débit de parole ne jouent pas directement un rôle sur la composition spectrale des voyelles et des consonnes. Selon lui, les indices spectraux restent relativement stables sous l'effet des variations temporelles. Il conclut que c'est essentiellement la vitesse des transitions formantiques qui varie en fonction du débit de parole. Harrington (2010) commente ce débat en soulignant que *"l'influence du débit sur l'espace vocalique n'est pas toujours claire, pas seulement parce que les locuteurs n'augmentent pas le débit par le même facteur, mais aussi parce qu'il pourrait y avoir des réorganisations articulatoires accompagnant les variations de débit."* [1]

Les études classiques sur les relations articulatoire / acoustique associées aux occlusives (Delattre *et al.*, 1955; Öhman, 1966; Kewley-Port, 1982) ont cherché à identifier des propriétés acoustiques invariantes pour décrire le lieu d'articulation. Le concept de locus formantique a été proposé par Delattre *et al.* (1955). Ce paramètre étant sensible aux contextes vocaliques (Öhman, 1966; Kweley-Port, 1982), il a fallu chercher des indices plus fiables qui prendraient en compte les aspects dynamiques de la production de la parole. Les équations de locus ont été proposées comme étant des invariants relationnels du lieu d'articulation (Sussman *et al.*, 1991; Sussman & Shore, 1996; Sussman *et al.*, 1998; Lindblöm & Sussman, 2012). Elles sont aussi de bons indicateurs du degré de coarticulation entre consonne et voyelle (Lindblöm, 1963; Krull, 1989; Duez, 1992; Fowler, 1994). Dans ces études, les équations de locus ont été mesurées indépendamment des phénomènes temporels que sont la longueur et la durée des voyelles ; à l'exception des études de Krull (1989) et Duez (1992) qui montrent un effet des durées sur les paramètres des équations de locus en comparant parole spontanée et parole contrôlée. Plus récemment, Berry & Weismer (2013) ont étudié le degré de coarticulation

1. Traduction française du paragraphe page 92.

dans différents débits de parole. Les résultats mettent en exergue que lorsque le débit augmente, la pente des équations de locus devient plus élevée tandis que l'ordonnée à l'origine baisse. Une pente proche de 1 indique une forte coarticulation alors qu'une pente proche de 0 révèle une absence de coarticulation. L'accélération du débit provoquerait donc un accroissement du chevauchement coarticulatoire.

Dans cet article, l'impact de l'opposition de longueur vocalique et des variations de débit de parole sur les propriétés spectrales de séquences occlusive-voyelle est évalué. De même, le degré de coarticulation et son influence sur les coefficients de l'équation de locus sont abordés.

2 Expérience I

2.1 Participants

Sept locuteurs jordaniens masculins âgés de 27 à 35 ans ont participé à cette expérience. L'un des locuteurs est le premier auteur. Ils sont originaires de 3 régions différentes de Jordanie (Ma'an, Amman et Jerash). Ils ne présentent aucun trouble du langage.

2.2 Stimuli

Les locuteurs devaient lire des séquences CVC enchâssées dans une phrase porteuse (/ħaka CVC hassa/ : « Il a dit CVC à l'instant »). Les séquences apparaissaient l'une après l'autre en alphabet arabe vocalisé (les diacritiques représentant les voyelles) sur un écran d'ordinateur.

Trois occlusives voisées et deux occlusives non-voisées ont été étudiées /b,d,g,t,k/. Chacune de ces occlusives apparaissait dans un mot de la langue ayant une structure CVC. Ces 5 occlusives cibles ont été combinées avec 6 voyelles : 3 courtes –{i, a, u}– et 3 longues –{iː, aː, uː}/ dans deux positions lexicales (initiale $C_{cible}VC$ vs. finale CVC_{cible}). L'autre consonne du mot était sélectionnée au hasard en fonction des mots existant dans le vocabulaire en arabe jordanien, soit un total de 80 mots différents. La fréquence des mots n'a pas été contrôlée.

2.3 Procédure

Il était demandé aux locuteurs de lire les séquences de la manière la plus naturelle possible sans chercher à parler correctement. Un programme écrit en `Python`[2] avec les librairies `Pygame`[3] et `pyfribidi`[4] contrôlait le cours de l'expérience (durées d'affichage, écriture de droite à gauche, ordre de présentation des séquences). Les séquences (mot en contexte avec la phrase porteuse) étaient présentées sur l'écran d'ordinateur dans un ordre aléatoire et à un rythme relativement soutenu mais confortable (une phrase toutes les 1500 ms). Les participants ne connaissaient pas les séquences avant le début de l'enregistrement. Chaque mot était répété 10 fois. Au final, 600 séquences ont été obtenues par locuteur (5 consonnes $\times$ 2 positions syllabiques $\times$ 6 voyelles $\times$ 10 répétitions).

Les séquences CVC ont été segmentées et transcrites à la main avec le logiciel Praat (Boersma & Weening, 2012). Les fréquences des 3 premiers formants de chaque séquence CV/VC cible et leurs coordonnées temporelles ont été extraites automatiquement grâce à un script Praat développé par les

2. http ://www.python.org
3. http ://www.pygame.org
4. http ://pyfribidi.sourceforge.net/

auteurs. L'algorithme d'extraction *Burg* (analyse LPC par auto-corrélation) a été employé avec une fenêtre d'analyse de 0.025s et un pas de 0.00625s.

Ces données ont été placées dans un fichier de résultats unique et ensuite traitées par un script R (R Core Team, 2012) chargé de procéder à l'analyse des données. Dans un premier temps, les tracés formantiques ont été lissés par moyennage glissant sur toutes les suites de 2 valeurs successives afin de compenser les variations trop rapides des fréquences formantiques (filtrage passe-bas). Les fréquences initiales / finales (en fonction de la position de l'occlusive étudiée dans le mot) et médianes des 3 premiers formants de chaque tracé formantique ont été sélectionnées et associées aux variables contrôlées dans l'expérience (occlusive cible, longueur vocalique, timbre vocalique, position de l'occlusive dans le mot, locuteur). Les positions de début et de fin de chaque voyelle, et la durée associée, ont été déterminées à partir du début de l'apparition des résonances formantiques des deux premiers formants en incluant les transitions formantiques *on-glide* et *off-glide*.

Des régressions linéaires ont ensuite été produites à partir de ces points afin de calculer les paramètres de chacune des 80 équations de locus (5 consonnes $\times$ 2 longueurs vocaliques $\times$ 2 positions syllabiques $\times$ 4 locuteurs). Les mesures, dont le résidu (l'erreur), entre la valeur observée dans l'espace $F2_{onset} \sim F2_{mid}$ et celle prédite par la régression était supérieur à la moyenne des valeurs absolues des résidus $+1.5$ fois l'écart-type, ont été identifiées comme *outlier* et retirées de l'analyse afin d'éviter l'impact de valeurs extrêmes sur les pentes des équations de locus. Suite à cette première phase d'analyse, une seconde phase de régressions linéaires a été produite sans ces valeurs. Ces données extrêmes représentent 7% des données [5]. Les données de position finale (VC) ne sont pas discutées ici.

2.4 Résultats

2.4.1 Durée des voyelles

Comme attendu, les voyelles longues ont bien des durées plus longues que les voyelles courtes. Ces données sont présentées dans la Table 1. Cet effet est significatif dans une ANOVA à deux facteurs à mesures répétées avec les locuteurs comme facteur aléatoire ($F_{(1,4)} = 57.92, p < 0.01$).

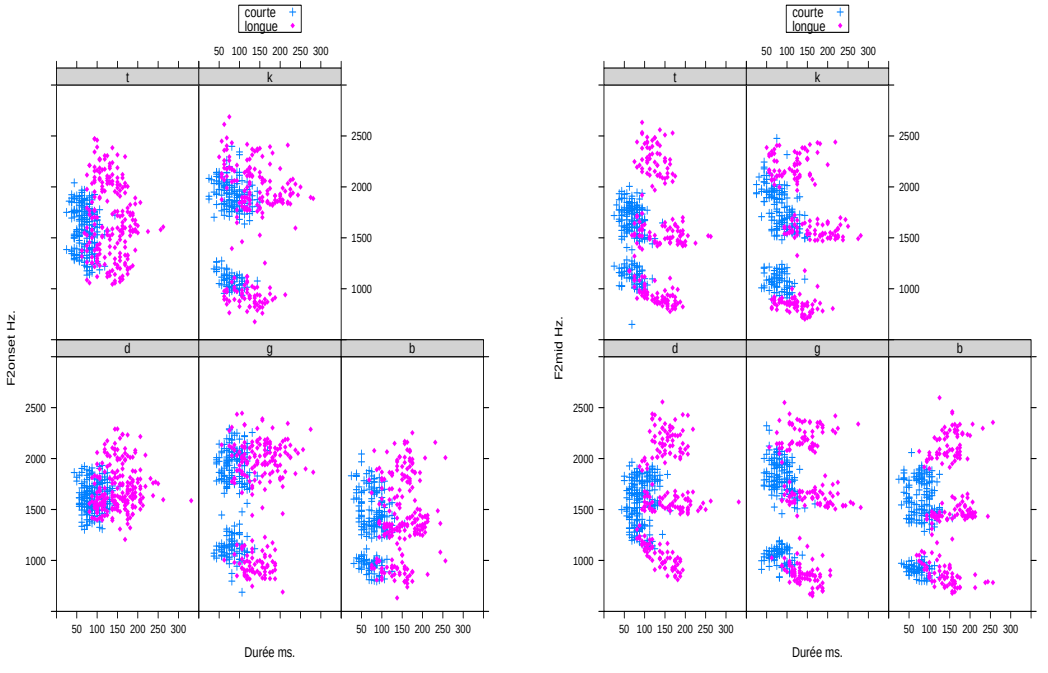
	i	a	u
courte	77 (25)	90 (24)	78 (21)
longue	135 (37)	162 (47)	134 (35)

TABLE 1 – Moyenne (et écart-type) des valeurs de la durée vocalique en ms. des voyelles longues comparées aux voyelles courtes.

2.4.2 Propriétés spectrales des formants

Les fréquences des trois premiers formants au début (F_{onset}) et au milieu (F_{mid}) de la voyelle ont été mesurées pour chaque séquence afin d'observer l'impact de la longueur vocalique sur la consonne et sur la voyelle. La Figure 1 illustre la distribution de ces changements de qualité spectrale (sur $F2$ uniquement) des consonnes ($F2_{onset}$) et des voyelles ($F2_{mid}$), en association avec les variations de durée vocalique, en fonction de la longueur phonologique de la voyelle. Dans les deux représentations de la Figure 1, la distribution horizontale des points indique la variation temporelle des voyelles au fil du temps et celle correspondant à l'axe vertical révèle leur répartition fréquentielle. La distribution

5. L'usage de méthodes de régression robuste est en cours afin d'adopter une meilleure approche des données extrêmes.



(a) (b)

FIGURE 1 – Illustration de la distribution des fréquences (axe vertical) de $F2_{onset}$ (a) et $F2_{mid}$ (b) et des durées de la voyelle (axe horizontal) pour chaque consonne en fonction de la longueur vocalique.

des voyelles longues ($F2_{mid}$ dans la Figure 1(b)) distingue 3 groupes qui correspondent aux 3 timbres vocaliques : {i,a,u}. Les voyelles longues sont plus dispersées sur l'axe horizontal que les voyelles courtes et elles sont facilement distinctes des voyelles courtes. Le même phénomène est remarqué sur l'axe vertical ; les fréquences des voyelles longues sont plus dispersées que celles des voyelles courtes. Ces dernières sont plus concentrées sur les deux axes que les voyelles longues. Cette concentration est plus importante dans la distribution des $F2_{onset}$ (dans la Figure 1(a)) que celle des $F2_{mid}$. Une tendance globale est observée (nettement plus claire en $F2_{mid}$) lorsque la durée des voyelles décroît, les points se dirigent vers une zone de fréquence centrale.

2.4.3 Coarticulation CV

Les équations de locus ont été mesurées pour chaque consonne dans chaque catégorie de longueur vocalique afin d'explorer l'impact des variations temporelles provoquées par l'opposition de longueur vocalique sur la coarticulation CV. La Table 2 présente les valeurs moyennes des pentes et des ordonnées à l'origine ainsi que les R^2 pour les 5 occlusives en contexte de voyelle courte vs. longue. Il apparaît que les consonnes produites avec des voyelles courtes ont des pentes relativement plus élevées et des ordonnées à l'origine globalement plus basses que lorsqu'elles sont produites en coarticulation avec une voyelle longue. Ces différences sont significatives pour la pente ($F_{(1,6)} = 9.74, p < 0.05$) et pour l'ordonnée à l'origine ($F_{(1,6)} = 20.62, p < 0.01$).

	Pente		Ordonnée à l'origine		R^2	
C	V courtes	V longues	V courtes	V longues	V courtes	V longues
b	0.77	0.69	244	363	0.94	0.90
d	0.54	0.39	774	1083	0.86	0.84
g	1.03	0.85	60	334	0.94	0.86
t	0.69	0.62	551	698	0.93	0.93
k	1.09	0.94	-49	210	0.94	0.93

TABLE 2 – Valeurs moyennes des pentes, des ordonnées à l'origine et des R^2 des équations de locus calculées pour chaque consonne coarticulée avec les voyelles courtes et les voyelles longues en position initiale (Expérience I).

3 Expérience II

3.1 Participants

Quatre locuteurs français masculins âgés de 22 à 28 ans ont participé à cette expérience. Ils étaient étudiants à l'Université de Nantes lors de l'enregistrement et ne présentent aucun trouble du langage.

3.2 Stimuli

Les locuteurs devaient lire des séquences CVC enchâssées dans une phrase porteuse (« Il a dit CVC huit fois »). Les phrases à prononcer apparaissaient sur un écran d'ordinateur à 3 débits différents en 3 blocs successifs (rapide, puis moyen et enfin lent). Une barre de progression s'affichait progressivement sous les phrases pour donner une indication aux participants du débit voulu, avec une vitesse « cible » de 70, 140 et 210 ms / syllabe respectivement pour les débits rapide, moyen et lent. Chaque mot était répété 7 fois dans un ordre aléatoire à l'intérieur du bloc.

Au total, 3 occlusives voisées et 3 occlusives non-voisées ont été examinées /b,d,g,p,t,k/. Chacune de ces occlusives apparaissait dans un mot de la langue ayant une structure CVC. Ces 6 occlusives cibles ont été combinées avec 4 voyelles {i, a, u, y} dans 2 positions lexicales (initiale $C_{cible}VC$ vs. finale CVC_{cible}). L'autre consonne du mot était sélectionnée au hasard en fonction des mots existant dans le vocabulaire français. Au total, 48 mots différents (produits à 3 débits) ont été prononcés. La fréquence des mots n'a pas été contrôlée. Les mêmes procédures de segmentation et d'extraction des données spectrales ont été employées que celles mises en œuvre dans l'expérience I.

3.3 Résultats

3.3.1 Durée des voyelles

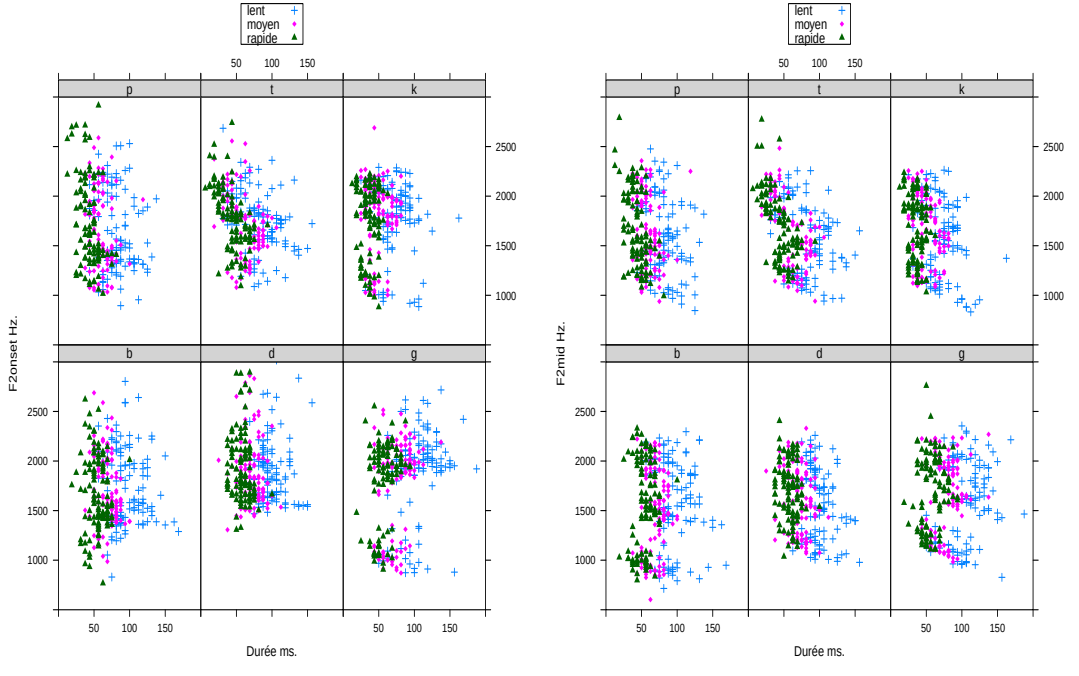
La Table 3 présente les durées moyennes et leur dispersion pour chaque condition de débit. L'effet global du débit sur la durée est significatif ($F_{(2,6)} = 12.85, p < 0.01$).

3.3.2 Propriétés spectrales des formants

L'analyse des données a été restreinte aux 3 voyelles cardinales {i, a, u} afin de pouvoir comparer plus nettement les données de l'arabe jordanien et du français (dans cette section uniquement).

	i	a	u	y
rapide	43 (13)	43 (12)	45 (10)	44(14)
moyen	58 (15)	80 (10)	62 (12)	62(18)
lent	82 (24)	105 (21)	87 (21)	80(27)

TABLE 3 – Moyenne (et écart-type) des durées vocaliques (en ms) pour les 3 débits.



(a) (b)

FIGURE 2 – Illustration de la distribution des fréquences (axe vertical) de $F2_{onset}$ –Fig. 2(a)– vs. $F2_{mid}$ –Fig. 2(b)– et des durées des voyelles /i,a,u/ (axe horizontal) pour chaque consonne en fonction du paramètre de débit (Expérience II).

Contrairement à la distribution des voyelles de l'arabe, celle des F2$_{mid}$ du français (Figure 2(b)) différencie peu les 3 groupes qui correspondent aux 3 voyelles : {i,a,u} . De plus, les 3 débits ne sont pas dissociables, ils présentent plutôt une répartition continue. Par comparaison avec les données de l'arabe, il apparaît que la distribution verticale des points est plus étendue pour F2$_{onset}$ et F2$_{mid}$ et celle horizontale moins étendue. La fréquence des voyelles tend très légèrement à se diriger vers une zone de fréquence centrale à débit rapide mais ce phénomène est nettement moins marqué que dans les données de l'arabe.

3.3.3 Coarticulation CV

Les équations de locus ont été calculées pour chaque consonne associée à chaque débit afin d'explorer l'impact du débit de parole sur la coarticulation CV. Les $F2_{onset}$ et $F2_{mid}$ ont été employés afin de calculer les paramètres des équations de locus. La Table 4 présente les valeurs moyennes des pentes, des ordonnées à l'origine et des R^2 pour les 6 occlusives produites aux 3 débits. Les pentes et les ordonnées à l'origine sont influencées par le débit : les consonnes produites avec un débit rapide ont

des pentes relativement plus élevées et des ordonnées à l'origine relativement moins élevées que celles des mêmes consonnes produites avec un débit lent. L'effet global du débit (3 niveaux) est significatif pour les pentes ($F_{(2,6)} = 5.595, p < 0.05$) et les ordonnées à l'origine ($F_{(2,6)} = 9.478, p < 0.05$).

	Pente			Ordonnée à l'origine			R^2		
C	rapide	moyen	lent	rapide	moyen	lent	rapide	moyen	lent
b	0.73	0.57	0.56	806	834	516	0.73	0.69	0.51
d	0.49	0.32	0.22	996	1295	1449	0.63	0.64	0.45
g	1.02	0.99	0.74	114	197	701	0.70	0.68	0.53
p	1.06	0.86	0.93	-75	277	139	0.80	0.62	0.76
t	0.81	0.75	0.57	412	526	851	0.92	0.73	0.73
k	1.03	0.84	0.83	45	422	646	0.93	0.72	0.78

TABLE 4 – Valeurs moyennes des pentes, des ordonnées à l'origine et des R^2 des équations de locus calculées pour chaque consonne en débit rapide, moyen et lent.

4 Discussion

L'impact de 2 types de variations temporelles (débit et longueur phonologique) sur les propriétés spectrales et coarticulatoires a été étudié dans 2 langues différentes. Ces 2 types modifient globalement la qualité de la voyelle et celle de la consonne à des niveaux différents. Dans les 2 langues, quand la durée décroît, les fréquences des voyelles et des consonnes se dirigent vers une zone de fréquence centrale. Ce phénomène est mieux visible en arabe et sur $F2_{mid}$. Ces résultats sont en accord avec les travaux de Lindblöm (1963) et de Berry & Weismer (2013).

La variation temporelle de la voyelle en arabe jordanien est notamment liée à l'opposition de durée phonologique. La qualité de la consonne semble être plus stable que celle de la voyelle ; autrement dit l'opposition voyelles longue vs. courte est moins marquée en $F2_{onset}$ qu'en $F2_{mid}$. Néanmoins, l'opposition de durée phonologique influence la relation coarticulatoire CV. Le chevauchement coarticulatoire entre la consonne et la voyelle est plus important pour les voyelles courtes que dans le cas des voyelles longues. Les résultats de cette étude mettent en relief que les pentes des voyelles courtes sont plus élevées et leurs ordonnées à l'origine sont plus basses que celles des voyelles longues.

En français, l'impact des variations temporelles liées au débit de parole paraît avoir un comportement similaire au précédent. Par contre, les qualités de la voyelle et de la consonne sont influencées d'une manière moins importante lorsque le débit de parole change. Par ailleurs, la relation CV est sensible aux variations de débit de parole : en fonction de l'augmentation du débit, la valeur de la pente s'accroît et la valeur des ordonnées à l'origine diminue.

Les fréquences des $F2_{onset}$ dans les deux langues ont un comportement similaire. La concentration de points en un nuage signifie que la consonne est moins coarticulée. Les alvéolaires /d,t/ semblent être relativement résistantes : les points sont concentrés dans une zone de fréquence globalement réduite pour /d/ et plus importante pour /t/. Quant aux vélaires /k,g/, elles forment 2 nuages de points désignant l'antériorité / postériorité des voyelles longues et courtes avec lesquelles elles sont produites. Les bilabiales /b,p/ ne semblent pas influencer les voyelles puisque la dispersion des points des $F2_{onet}$ est aussi importante que celle des $F2_{mid}$. Cette différence de coarticulation est démontrée par les valeurs de pente et d'ordonnée à l'origine pour chaque consonne.

Références

BERRY J. & WEISMER G. (2013). Speaking rate effects on locus equation slope. *Journal of Phonetics*, **41**, 468–478.

BOERSMA P. & WEENING D. (2012). Praat : Doing phonetics by computer. Computer program.

DELATTRE P., LIBERMAN A. & COOPER F. (1955). Acoustical loci and transitional cues for consonants. *The Journal of the Acoustic Society of America*, **27**(4), 769–773.

DUEZ D. (1992). Second formant locus-nucleus patterns : An investigation of spontaneous french speech. *Speech Communication*, **11**(4-5), 471–427.

FOWLER C. (1994). Invariants, specifiers, cues : An investigation of locus equations as information for place of articulation. *Perception and Psychophysics*, **55**, 597–610.

GAY T. (1978). Effect of speaking rate on vowel formant movements. *Journal of the Acoustical Society of America*, **63**, 223–30.

HADDING-KOCH K. & ABRAMSON A. S. (1964). Duration versus spectrum in swedish vowels : Some perceptual experiments. *Studia Linguistica*, **18**(2), 94–107.

HARRINGTON J. (2010). *Acoustic Phonetics*, In *The Handbook of Phonetic Sciences*, p. 81–129. Blackwell Publishing Ltd.

KEWLEY-PORT D. (1982). Measurement of formant transitions in naturally produced stop consonant-vowel syllables. *Journal of the Acoustical Society of America*, **72**(2), 379–389.

KRULL D. (1989). Second formant locus patterns and consonant-vowel coarticulation in spontaneous speech. *Phonetic experimental research at the institute of linguistics university of Stockholm-PERILUS*, **10**, 87–108.

KWELEY-PORT D. (1982). Measurement of formant transitions in naturally produced stop consonant-vowel syllables. *Journal of the Acoustical Society of America*.

LINDBLÖM B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, **35**(11), 1773–1781.

LINDBLÖM B. & SUSSMAN H. M. (2012). Dissecting coarticulation : How locus equations happen. *Journal of Phonetics*, **40**(1), 1–19.

ÖHMAN S. (1966). Coarticulation in VCV utterances : Spectrographic measurements. *The Journal of the Acoustical Society of America*, **39**(1), 151–168.

O'SHAUGHNESSY D. (1986). The effects of speaking rate on formant transitions in French synthesis-by-rule. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86.*, volume 11, p. 2027–2030.

R CORE TEAM (2012). *R : A Language and Environment for Statistical Computing*. Vienna, Austria : R Foundation for Statistical Computing. ISBN 3-900051-07-0.

SUSSMAN H. M., FRUCHTER D., HILBERT J. & SIROSH J. (1998). Linear correlates in the speech signal : The orderly output constraint. *Behavioral and Brain Siences*, **21**, 241–299.

SUSSMAN H. M., MCCAFFREY H. A. & MATTHEWS S. A. (1991). An investigation of locus equations as a source of relational invariance for stop consonant place categorization. *Journal of the Acoustical Society of America*, **90**, 1309–1325.

SUSSMAN H. M. & SHORE J. (1996). Locus equation as phonetic descriptors of consonantal place of articulation. *Psychonomic Society, Inc.*, **58**(6), 936–946.

TSUKADA K. (2009). An acoustic comparison of vowel length contrasts in Arabic, Japanese and Thaï : Durational and spectral. *International Journal on Asian Language Processing*, **19**(4), 127–138.

Incidence de la chirurgie naso-sinusienne sur la qualité vocale : étude d'un cas clinique

Lise Crevier-Buchman[1, 2] Angélique Amelot[1] Bénédicte Mas[2] Mathilde Giron[2] Pierre Bonfils[2,3]

(1) Laboratoire de Phonétique et Phonologie, UMR7018, Univ. Paris3 Sorbonne Nouvelle, 19 rue des Bernardins 75005 Paris, France

(2) Hôpital Européen Georges Pompidou, Univ. Paris5, 20 rue Leblanc 75015 Paris

(3) Cognition and Action Group, CNRS MD 8257, SSA and University Paris 5, Paris

lise.buchman@numericable.fr, angelique.amelot@univ-paris3.fr, benedictemas@yahoo.fr, mathilde.giron@orange.fr, pierre.bonfils@aphp.fr

RESUME

Les fosses nasales participent à la résonance vocale et toute modification de ces structures peut altérer la qualité vocale. Le rôle des sinus comme résonateurs dans la production vocale reste plus controversé. Le but de notre étude prospective était d'explorer d'éventuelles modifications acoustiques chez un chanteur professionnel en pré et post-opératoire après chirurgie naso-sinusienne unilatérale. A partir de la lecture d'un texte, nous avons extrait les voyelles /a,i,u/ pour mesurer les paramètres acoustiques de fréquence (F0), des formants F1 et F2, de leur largeur de bande, et de qualité vocale (LTAS et H1*-H2*). L'étude a été complétée par une auto-évaluation de la qualité de voix. Nos résultats n'ont pas permis de mettre en évidence de différence statistiquement significative des paramètres acoustiques bien que le patient ait signalé une impression d'amélioration vocale chantée. Ces résultats pour le français confirment ceux de la littérature et peuvent servir à informer les patients.

ABSTRACT

Impact of Sinus Surgery on Voice Quality: Case Study

The nasal cavity contributes to voice resonance and changes in these structures can alter voice quality. The role of the sinus as resonators for voice production is still controversial. The aim of our prospective study was to evaluate acoustic changes for a professional singer before and six month after sinus surgery. We extracted the vowels /a,i,u/ from a French text and a self evaluation of voice handicap. Objective measures of fundamental frequency, formant frequencies (F1 & F2) and their bandwidth, and voice quality parameters (LTAS and H1*-H2*) were performed. No statistical differences were identified for all our measures although the patient felt an improvement in his singing voice. Our results for French language confirm what has been observed in most of the international literature in other languages. Therefore, we can provide an informed consent based on objective measures for patients undergoing sinus surgery.

MOTS-CLES : Sinus, chirurgie naso-sinusienne, qualité vocale, paramètres acoustiques, fréquence fondamentale, formants

KEYWORDS: Paranasal sinuses, sinus surgery, voice quality, acoustic parameters, fundamental frequency, formant frequency

1 Introduction

Les résonateurs du système phonatoire correspondent à l'ensemble des cavités supra-glottiques et supra-laryngées. Le son de base émis par la vibration des plis vocaux est modulé par le conduit vocal, constitué par le pharynx, la cavité buccale, la cavité nasale et la cavité labiale. Cet ensemble de résonateurs couplés se comporte alors comme un filtre acoustique complexe, en modifiant la balance spectrale du son source, et lui conférant ainsi un timbre particulier. Les cavités nasales ou fosses nasales sont constituées d'une charpente osseuse complexe recouverte d'une muqueuse. Grâce à de petits orifices appelés ostia, les fosses nasales communiquent avec les sinus paranasaux qui se drainent dans les fosses nasales. Les quatre sinus paranasaux (frontal, ethmoïdal, maxillaire et sphénoïdal) sont des cavités remplies d'air et creusées à l'intérieur des os du crâne. Si la description anatomique et physiologique des sinus a rapidement été établie (Galen, 13&201 A.D.), leur rôle et la raison de leur présence sont plus controversés. En ce qui concerne le rôle des sinus paranasaux dans la parole, plusieurs auteurs leur admettent un rôle de résonateur dès le 17ème siècle (Haarwood, 1799 : Voltine 1888). Une chirurgie sinusienne n'a pas seulement un impact sur les sinus mais sur la cavité nasale dans son ensemble car elle permet, selon les techniques, d'agrandir la fosse nasale et de créer un espace de résonnance plus grand. Actuellement, la chirurgie sinusienne est proposée aux patients atteints de rhinosinusite chronique, de polypose nasale, d'obstruction nasale et de pathologies tumorales des sinus. Les résultats de la chirurgie sur l'obstruction nasale et les infections chroniques sont très satisfaisants ; elle permet aussi une récupération de l'odorat et une meilleure ventilation ; mais les modifications acoustiques possibles ne sont pas bien établies.

La littérature concernant les effets d'une telle chirurgie sur la qualité vocale reste pauvre. Une étude menée par Hösemann, *et al.*, (1998) sur 21 germanophones, a montré que des modifications vocales sont possibles après une telle chirurgie et qu'il est nécessaire d'en informer les patients. Cependant, les auteurs soulignent la diversité des effets acoustiques de la chirurgie sinusienne et restent prudents quant à ces résultats étant donné la petite cohorte. Une étude plus récente de Acar *et al.* (2013) indique qu'il n'existe pas de différence significative pour quatre paramètres acoustiques (HNR, F0, Jitter et Shimmer) après ablation de polypes dans les sinus paranasaux. Certains auteurs constatent néanmoins l'existence d'un effet de la chirurgie naso-sinusienne sur les voyelles nasalisées (Chen *et al*., 1997 ; Hosemann *et al*., 1998) et les consonnes nasales (Chen *et al*., 1997 ; Kim *et al*., 2014) par des mesures de nasalance (mesures du ratio entre énergie acoustique nasale et orale). Certains articles font état du caractère subjectif de la perception d'un changement de la qualité vocale après chirurgie naso-sinusienne (Chen *et al*., 1997 ; Hosemann *et al*., 1998 ; Sonegheta *et al*. 2002). L'utilisation des questionnaires d'auto-évaluation de la qualité de voix a montré que 20% des patients signalaient un changement positif de qualité vocale après cette chirurgie. Les auteurs soulignent la nécessité d'en informer les patients, notamment les professionnels de la voix. »

L'imprécision quant aux résultats vocaux de cette chirurgie chez des professionnels de la voix nécessite une investigation approfondie, pour pouvoir répondre de façon objective à la demande d'information et aux inquiétudes sur un risque éventuel de modification post-opératoire du son de leur voix. Il n'existe pas à notre connaissance d'études qui se sont intéressées à la modification possible de la qualité de la voix chez les locuteurs francophones après chirurgie sinusienne. Or le français a la particularité d'avoir une opposition fine en ce qui concerne les voyelles nasales et les voyelles nasalisées. Dans la mesure où la chirurgie naso-sinusienne modifie le volume des cavités nasales, on peut supposer une modification du couplage acoustique des sinus comme un résonateur d'Helmholtz, et des propriétés acoustiques de la muqueuse des voies aériennes supérieures. Nous

avons cherché à objectiver une éventuelle modification des paramètres acoustiques et de la qualité vocale chez un chanteur lyrique professionnel en préopératoire et trois mois post-opératoire.

2 Matériel et Méthodes

2.1 Patient

Un homme de 51 ans, chanteur lyrique professionnel, ténor, a présenté une tumeur bénigne de type papillome inversé dans le sinus sphénoïdal gauche. L'ablation chirurgicale a été décidée en avertissant le patient des risques vocaux potentiels mais non connus et non documentés dans la littérature. Le risque était en rapport avec une modification du volume des cavités nasales et sinusiennes et par conséquent des résonances nasales. Le patient a signé un consentement éclairé. L'intervention a consisté en une ouverture large du sinus ethmoïdal et sphénoïdal gauche et d'une turbinectomie moyenne gauche pour la voie d'abord endonasale. L'opération a été réalisée par navigation assistée par ordinateur le 15 janvier 2015.

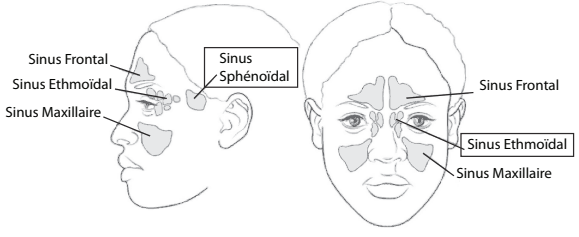


FIGURE 1 : schéma de la place anatomique des 4 paires de sinus de la face

Les bilans pré et post-opératoire à 6 mois se sont déroulées dans les mêmes conditions, en consultation d'oto-rhino-laryngologie et phoniatrie de l'Hôpital Européen Georges Pompidou. En pré-opératoire, le patient se plaignait d'une obstruction nasale chronique avec rhinite et sécrétions qui coulaient dans le pharynx. L'examen ORL vélo-pharyngo-laryngé a été réalisé en naso-fibroscopie souple (Kay-Pentax FNL10RP3) équipé d'une lumière froide (halogène CLK-4, Olympus) dont l'extrémité est reliée à un système d'enregistrement vidéo (DigitalStrobe, RLS91000, Kay Elemetrics) à une fréquence de 25 im/sec et une résolution de 568 × 454 pixels.

2.2 Enregistrements, corpus

L'enregistrement acoustique pré et post-opératoire de la voix a été réalisé avec la station d'enregistrement Kay Elemetrics CSL (Computer Speech Lab) 4300 et un microphone casque AKG C520 dans une pièce acoustiquement calme.

Le corpus consistait en la tenue de la voyelle /a/ à une hauteur et intensité confortable, pour l'analyse acoustique des paramètres de fréquence et la lecture du texte « la bise et le soleil ».

Le patient a rempli un questionnaire d'auto-évaluation sur sa qualité de voix : le Voice Handicap Index (VHI30) (Jacobson *et al*., 1997). Ce questionnaire est composé de 30 items comprenant 3 sous-échelles fonctionnelle, physique et émotionnelle avec 5 réponses possibles : « jamais », « presque jamais », « parfois », « presque toujours », « toujours », et une cotation globale du ressenti sur sa voix « moins bonne », « pareille», « meilleure » qu'avant.

2.3 Mesures acoustiques

2.3.1 Voyelle isolée

Les mesures acoustiques ont été réalisées à partir de la voyelle /a/ tenue pendant 3 secondes pour extraire la F0 et les paramètres d'instabilité instantanée (Jitter et Shimmer) ainsi que le rapport Bruit sur Harmoniques (NHR) avec Multi-Dimensional Voice Program (MDVP) (Version 3.3, Kay Elemetrics Corporation Lincoln Park, New Jersey, USA).

2.3.2 Lecture

Les mesures de durées, des deux premiers formants, la largeur de bande et la valeur de F0 ont été extraites au milieu de chaque voyelle /i a u/ du texte lu. La mesure de H1*-H2* a été effectuée sur la voyelle /a/ (Hansen, 1996). Le spectre moyenné à long terme (LTAS), qui permet de rendre compte des résonances supra-glottiques (Kitzing *et al*., 1993 ; Miller *et al*., 2002) a été fait mesuré sur les segments voisés du texte. Ces mesures ont été réalisées avec Praat (Boersma *et al*., 2016).

3 Résultats

3.1 L'examen laryngé

En préopératoire, l'examen laryngé est sans particularité en dehors d'une inflammation diffuse liée à la rhinorrhée postérieure. En phonation on observe un léger comportement d'effort avec fermeture serrée du plan glottique et participation des bandes ventriculaires. L'ondulation muqueuse des plis vocaux est bien conservée, bilatérale et symétrique. En post-opératoire le patient signale une amélioration du timbre de sa voix qu'il qualifie de plus claire, et une plus grande aisance dans les aigus de sa voix chantée. L'examen laryngé montre une diminution du serrage supra-glottique au niveau des bandes ventriculaires en phonation et une disparition des sécrétions et de l'inflammation.

3.2 Les tests d'auto-évaluation VHI

L'échelle VHI n'a pas permis de relever de plaintes vocales (score a zéro en pré et post-opératoire). A la question subsidiaire « considérez vous votre voix comme moins bonne, pareille ou meilleure qu'avant », le patient signale qu'il considère sa voix comme « meilleure » en post-opératoire.

3.3 Les analyses acoustiques

3.3.1 Paramètres acoustiques

3.3.1.1 Voyelle isolée

- Paramètres acoustiques mesurés sur la voyelle tenue /a/ pré post-op : F0= 204/182 Hz ; ds= 2,27/1,36 ; J= 0,29/0,47 ; S= 2,27/1,60 ; NHR= 0,108/0,098

Ces paramètres sont dans les limites de la normale compte tenu de la hauteur de la F0 sur cette tâche de voyelle tenue qu'il a produite en voix aigue.

3.3.1.2 Lecture

La durée moyenne pour les trois voyelles /i a u/ extraites des lectures est de 90 ms en pré-opératoire et de 92 ms en post-opératoire. Une analyse statistique t-test sur les deux groupes montre qu'il n'existe pas de différence significative de durée en pré- et post-opératoire sur les 3 voyelles (t= 0.30114, df = 122.95, p-value = 0.7638).

Avec une moyenne de 156 Hz en pré et en post-opératoire pour les trois voyelles confondues, une analyse statistique t-test ne montre pas de différence significative de F0 entre les deux enregistrements (t = 0.018676, df = 126.93, p-value = 0.9851).

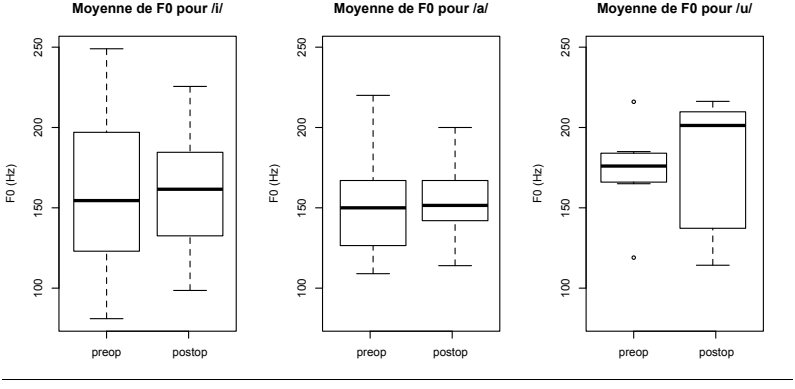


FIGURE 2 : représentation des moyennes de F0 pour les trois voyelles en pré et post-opératoire

Une analyse statistique anova à deux facteurs en fonction du temps de l'enregistrement (pré et post-opération) et en fonction des différentes voyelles montre qu'il n'existe pas non plus de différence significative (F(2,126) = 0.289, p=0.7495) de F0.

- Analyse formantique et largeur de bande pour F1 :

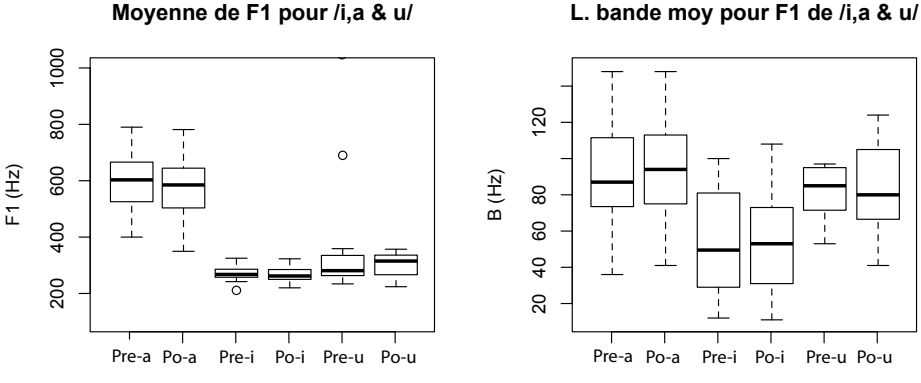


FIGURE 3 : valeur des formants F1 et largeur de bande pour les 3 voyelles aux 2 temps opératoires.

Une analyse statistique anova à deux facteurs en fonction du temps de l'enregistrement (pré- et post-opération) et en fonction des différentes voyelles montre qu'il n'existe pas de différence

significative ($F(2,126) = 0.566$, $p=0.569$) en ce qui concerne les valeurs de F1 ni pour les valeurs de la largeur de bande de F1 ($F(2,126) = 0.047$, $p= 0.955$).

- Analyse formantique et largueur de bande pour F2 :

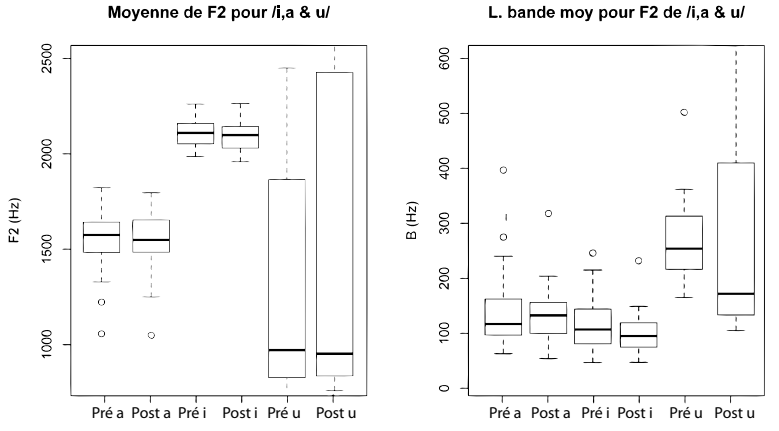


FIGURE 4 : valeur des formants F2 et largeur de bande pour les 3 voyelles aux 2 temps opératoires

Une analyse statistique anova en fonction du temps de l'enregistrement (pré- et post-opération) et en fonction des différentes voyelles montre qu'il n'existe pas de différence significative ($F(2,126) = 1.108$, $p= 0.333$) en ce qui concerne les valeurs de F2 tout comme les valeurs de la largeur de bande de F2 ($F(2,126) = 0.261$, $p= 0.771$).

- Différence entre les premier et deuxième harmoniques pour explorer la qualité vocale : H1* – H2*

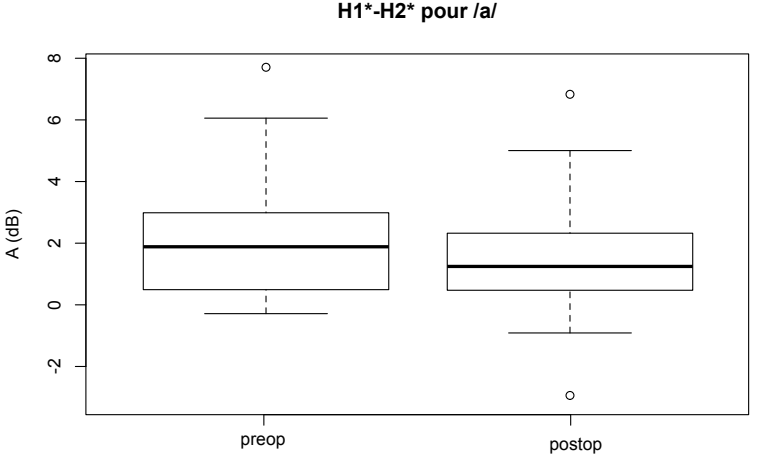


FIGURE 5 : valeur de la différence entre le 1er et le 2ème harmonique pour la voyelle /a/ aux 2 temps

Une analyse statistique anova en fonction du temps de l'enregistrement (pré- et post-opération) pour la voyelle /a/ montre qu'il n'existe pas de différence significative ($F(1,75) = 2.42$, $p= 0.124$) en ce

qui concerne les valeurs de H1*-H2*. Cependant, en post-opératoire, la différance diminue ce qui irait dans le sens d'une légère amélioration du timbre vocal qui peut être perçu par le patient.

- Spectre à long terme LTAS mesuré à partir de Praat.

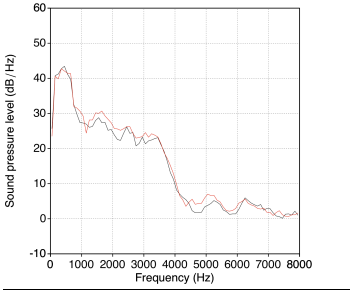


FIGURE 5 : superposition des spectres à long terme, en pré-op. (noir) et post-op. (rouge)

Ici nous avons exploré la répartition des énergies et des fréquences extraites des segments voisés du texte lu. Les 2 courbes sont superposées, la répartition d'énergie spectrale est semblable aux 2 temps.

4 Discussion et Conclusion

Le rôle des sinus de la face a été longuement discuté et n'est pas encore clairement connu (Eloy, 2005). Une revue de littérature menée par Keir (2007) interroge les fonctions possibles des sinus et conclue que la fonction principale des sinus serait d'améliorer les fonctions physiologiques du nez mais aussi avoir un rôle dans la résonance vocale. La résonance du spectre vocal est très intimement liée aux caractéristiques de la cavité nasale avec cependant une variabilité inter-individuelle liée aux différences morphologiques très importante des fosses nasales et des sinus. (Serrurier, Badin, 2008) Notre patient chanteur professionnel devait subir une chirurgie naso-sinusienne et nous avions besoin, d'un point de vue médico-légal, d'évaluer le retentissement de cette chirurgie sur sa qualité vocale. La région vélo-pharyngée n'ayant pas été modifiée, on pouvait s'attendre à une absence de modification des paramètres aérodynamiques. Notre étude a porté sur un cas avec une analyse de sa voix parlée et non chantée.

4.1 Relation entre paramètres acoustiques et perception d'une dysphonie

Une analyse acoustique (F0, intensité, NHR) et vidéostroboscopique sur 10 patients porteurs de sinusite chronique et 9 patients « sains » menée par Cecil *et al*. (2000) ont montré une absence de relation entre la dysphonie et la sinusite chronique. Les mêmes observations sont retrouvées chez Acar *et al*. (2013). Par contre, l'évaluation perceptive et acoustique réalisée par Chen et Metson (1997) montrent des changements acoustiques significatifs sur les consonnes nasales et les voyelles nasalisées corrélés aux jugements perceptifs. Nous n'avons pas exploré les phénomènes aérodynamiques ni les modifications segmentales qui auraient pu être en lien avec une nasalisation. Les voyelles non nasalisées n'étaient pas affectées et l'échelle VHI pour la voix parlée n'a pas permis d'objectiver de modification de qualité vocale. Dans le cas de notre chanteur, il aurait été intéressant d'utiliser le singing VHI (Morsomme *et al*. 2007). De plus, une exploration plus

spécifique avec phonétogramme pourrait apporté des informations sur la dynamique vocale dans tout le spectre des fréquences.

4.2 Relation entre Formants et qualité vocale

De nombreux auteurs s'accordent pour dire que les conséquences spectrales de la nasalité se situent surtout dans les basses fréquences, avec un aplatissement des pics spectraux autour de F1 et F2 (Maeda, 1982; Vaissière, 1995). Les résultats de l'étude acoustique de Hosemann *et al*. (1998) sur les voyelles /a, i, u/ nasalisées chez 21 patients opérés pour sinusite chronique ont été très hétérogènes : la largeur de bande des formants diminue après la chirurgie sur la voyelle [a] et augmente pour le [i]. Il n'y a pas de changements observés dans les chirurgies unilatérales. Chez les patients avec des polyposes massives, aucun changement ne fut observé sur les voyelles /a/ et /i/, tandis qu'une augmentation de la largeur de bande de la voyelle /u/ fut attestée.

Acar et al. (2014) ont enregistrés 43 patients turcs opérés des sinus pour polypose naso-sinusienne (PNS). L'analyse acoustique (mesure de la F0, jitter, shimmer et NHR) n'a pas retrouvé de changements statistiquement significatifs dans la qualité de la voix après l'opération. De même, l'étude menée par Brandt et al. (2014) sur 15 patients anglophones avec obstruction nasale, n'a pas démontré de changements statistiquement significatifs de la qualité vocale dans l'évaluation perceptive (auto-évaluation et évaluation par des auditeurs naïfs) et acoustique après la chirurgie. La plupart de ces études concluent que les patients, et notamment les professionnels de la voix, doivent être informés des altérations possibles de la parole après une intervention majeure sur les sinus (Acar *et al*., 2014 ; Hosemann *et al*., 1998).

Chez notre patient, on ne constate pas de modification de la hauteur ni de l'amplitude des deux premiers formants ce qui irait dans le sens d'une absence de nasalisation en post-opératoire et plus généralement une absence de modification du timbre vocalique. Il faut aussi noter que l'intervention était unilatérale ce qui pourrait expliquer l'absence d'expression d'un retentissement de la chirurgie naso-sinusienne comme l'a souligné Hosemann *et al*., (1998). Enfin, l'amélioration post-opératoire de l'état de la muqueuse pharyngo-laryngée, la disparition des sécrétions et des infections rhino-sinusiennes a certainement contribué à l'amélioration de la qualité vocale et au ressenti du patient.

Une étude objective plus complète sur les différents paramètres vocaux notamment avec des mesures aérodynamiques et de nasalance, ainsi qu'une étude perceptive, nous semblent nécessaire pour avoir des données plus approfondies sur le français, et pour pouvoir informer le patient sur des bases physiologiques et objectives, des risques de modification de sa qualité vocale.

Remerciements

Ce travail a bénéficié d'une aide du LabEx *EFL* (ANR-10-LABX-0083).

Références

ACAR A., CAYONU M., OZMAN M., ERYILMAZ A. (2014) Changes in Acoustic Parameters of Voice After Endoscopic Sinus Surgery in Patients with Nasal Polyposis *Indian J Otolaryngol Head Neck Surg* 66 (4):381–385

BOERSMA P., & WEENINK D.(2016). Praat: doing phonetics by computer [Computer program]. Version 6.0.13, retrieved 31 January 2016 from http://www.praat.org/

BRANDT MG., ROTENBERG BW., MOORE CC., BORNBAUM CC., DZIOBA A., GLICKSMAN JT., DOYLE PC. (2014) Impact of nasal surgery on speech resonance. *Ann Otol Rhinol Laryngol*. 123(8):564-70

CECIL M., TINDALL L., HAYTON R. (2001) The relationship between dysphonia and sinusitis: a pilot study. *Journal of Voice* 15 (2) :270-277

CHEN MY., METSON R. (1997) Effects of sinus surgery on speech *Arch Otolaryngology Head & Neck Surgery 123(8) 845-852*

ELOY P., NOLLEVAUX M-C., BERTRAND B. (2005) Physiologie des sinus paranasaux. *EMC-Oto-rhino-laryngologie* 2 20-416-A-10 : 185–197

GALEN 130-201 A.D. De Usu Partium Ix, 2 et seq. (Kuhn) 111, p. 691. (cité par BLANTON, Patricia L., et Norman L. BIGGS. « Eighteen hundred years of controversy: The paranasal sinuses », *American Journal of Anatomy* 124, nº 2 (février 1969): 135.)

HAARWOOD, 1799 & VOLTINE 1888 (cité par Blanton, Patricia L., et Norman L. Biggs. « Eighteen hundred years of controversy: The paranasal sinuses », *American Journal of Anatomy* 124, nº 2 (février 1969): 135.

HANSEN H. (1996). Glottal characteristics of female speakers : Acoustic correlates. J. Acoust. Soc. AM., 101(1) :466-481.

HOSEMANN W., GÖDE U., DUNKER J.E., EYSCHOLDT U. (1998) Influence of endoscopic sinus surgery on voice quality. *Eur Arch Otorhinolaryngology* 255 (10) 499-503

JACOBSON B.H., JOHNSON A., GRYWALSKI C., SILBERGLEIT A., JACOBSON G., BENNINGER M.S., ET NEWMAN C.W. « The Voice Handicap Index (VHI)Development and Validation ». *American Journal of Speech-Language Pathology* 6, nº 3 (1 août 1997): 66-70.

KEIR, J. « Why Do We Have Paranasal Sinuses? » *The Journal of Laryngology and Otology* 123, nº 1 (janvier 2009): 4-8. doi:10.1017/S0022215108003976

KIM S.D., PARK HJ., KIM GH., WANGSG., CHO KS. (2014). Changes and recovery of voice quality after sinonasal surgery. *Eur Arch Otorhinolaryngol*, 272(10) : 2853-9

KITZING P., AKERLUND L. (1993) Long Terme average spectrograms of dysphonic voices before and after theramy. *Folia Phoniatr*. 45 : 53-61

MAEDA S. (1982) The role of the sinus cavities in the production of nasal vowels. Acoustics, Speech and Signal Processing, IEEE International Conference on ICASSP 82

MILLER DG., SVEC J., SCHUTTE H.K. (2002) Measurement of characteristic leap interval between chest and falsetto registers. *Journal of Voice* ; 16 :8-19

MORSOMME D., GASPAR M., JAMART J., REMACLE M., VERDUYCKT I. (2007) Adaptation du Voice Handicap Index à la voix chantée. *Revue de Laryngologie-Ototlogie-Rhinologie*, 128 (5) :305-14

SERRURIER, A., BADIN P. (2008) A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data. *Journal of the Acoustical Society of America*, 123 (4): 2335-2355.

SONEGHETA R., PAULA SANTOSB R., BEHLAUC M., HABERMANND W., FRIEDRICHD G., STAMMBERGERD H. (2002) Nasalance Changes After Functional Endoscopic Sinus Surgery. *Journal of Voice*; 16(3):392-7.

VAISSIERE J. (1995) Nasalité et Phonétique In *Le voile du palais et la parole* In Colloque sur le voile pathologique. Lyon ; Société Française d'Acoustique

Influence de la quantité de données sur une tâche de segmentation de phones fondée sur les réseaux de neurones

Céline MANENTI Thomas PELLEGRINI Julien PINQUIER
IRIT, Université de Toulouse, UPS, Toulouse, France
{celine.manenti, thomas.pellegrini, julien.pinquier}@irit.fr

RÉSUMÉ

Dans cet article, nous décrivons une étude expérimentale de segmentation de parole en unités acoustiques sous-lexicales (phones) à l'aide de réseaux de neurones. Sur le corpus de parole spontanée d'anglais américain BUCKEYE, une F-mesure de 68% a été obtenue à l'aide d'un réseau convolutif, en considérant une marge d'erreur de 10 ms. Cette performance est supérieure à celle d'un annotateur manuel, l'accord inter-annotateurs étant de 62%. Restreindre les données d'apprentissage à celles d'un unique locuteur, 30 minutes environ, a eu pour conséquence moins de 10% de perte et utiliser celles de 5 locuteurs a permis d'atteindre des résultats similaires à utiliser plus de données. Utiliser le modèle entraîné avec le corpus anglais sur un petit corpus d'une langue peu dotée a donné des résultats comparables à estimer un modèle avec des données de cette langue.

ABSTRACT

Phone-level speech segmentation with neural networks : influence of the amount of data

In this article, we describe speech segmentation experiments at phone level with neural networks, on a U.S. English speech corpus. Using filter banks and a ConvNet, a 68% F-measure was obtained, with an error margin of 10 ms, a figure close to the annotation agreement rate between human annotators. We then studied the impact of reducing the training data size : the decrease in performance was less than 10% only, when training with data from a single speaker, 30 min of speech, instead of data from 5 speakers. More data did not bring further improvements. Finally, we used a CNN trained on U.S. English to segment a small corpus in Xitsonga, a less-resourced language from South Africa. The English model led to a similar performance when compared to a model trained on a subset of the Xitsonga data.

MOTS-CLÉS : Réseaux de neurones, phonèmes, segmentation, langues peu dotées.

KEYWORDS: Neural Networks, phonemes, segmentation, under-resourced languages.

1 Introduction

La segmentation de parole est le processus, humain (cognitif) ou automatique (quand il est réalisé par une machine), qui vise à identifier des frontières entre des unités (mots, syllabes, phonèmes) dans un enregistrement ou un flux de parole. En traitement automatique de la parole, c'est un sous-problème qui a diverses applications en reconnaissance automatique de la parole (RAP). Actuellement, la recherche automatique de segments permettant d'identifier des mots ou des unités sous-lexicales est portée par l'intérêt pour l'apprentissage non-supervisé de ces unités, soit pour construire un lexique

de prononciation en identifiant les mots et l'inventaire de phones sans connaissance linguistique *a priori* (Lee *et al.*, 2015), soit pour faire des liens avec l'humain et l'acquisition du langage, en particulier par les enfants (Jansen *et al.*, 2013).

Dans ce contexte, nous pouvons mentionner l'intérêt croissant de la communauté scientifique pour le traitement automatique de langues dites peu-dotées, avec l'organisation de conférences et de sessions spéciales dédiées à ce thème chaque année, comme par exemple le Workshop sur les technologies de la parole pour les langues peu-dotées *SLTU*. À celles-ci s'ajoutent des défis, comme le *Zero Resource Speech Challenge* (Versteegh *et al.*, 2015), qui consistait à identifier des mots ou pseudo-mots, et des unités sous-lexicales à partir d'enregistrements sonores uniquement. Les données utilisées dans ce défi étaient le corpus de parole spontanée BUCKEYE, d'anglais américain, et également un petit corpus d'une langue peu dotée, le Xitsonga, une langue d'Afrique du Sud.

Les réseaux de neurones profonds (DNN) sont devenus populaires dans le monde du traitement du signal en raison de leurs excellentes performances, en particulier en RAP. Selon le problème considéré, ils donnent des résultats similaires ou supérieurs aux GMM. Par exemple, Joshi *et al.* (2015) obtiennent un gain absolu de 3% en classification de voyelles. Les réseaux de neurones ont la particularité de pouvoir s'adapter aux données et à la tâche demandée, s'approchant de la forme la plus adaptée au problème. Bhargava & Rose (2015) ont ainsi constaté que le réseau pouvait imiter des représentations proches des bancs de filtres lorsqu'ils prenaient directement des fenêtres du signal temporel en entrée.

Dans ce travail, nous avons abordé la segmentation automatique en phones en modélisant les frontières de segments plutôt que les segments eux-mêmes. Nous comparons les performances, sur le corpus BUCKEYE, obtenues par des réseaux à couches cachées denses (*Multilayer Perceptron, MLP*) et par des réseaux convolutifs (*Convolutional Neural Networks, CNN*). Après une brève description de notre système dans la section 2, des corpus et métriques d'évaluation en section 3, nous comparons dans la section 4 différentes configurations des modèles (nombre de neurones, de filtres), et illustrons l'influence des données utilisées (faible quantité, langue différente) lors de l'apprentissage des réseaux de neurones.

2 Description du système

Le schéma 1 représente les différentes étapes de notre système permettant d'obtenir la segmentation en phones du signal de la parole. Les trois étapes sont détaillées dans les sous-sections suivantes.

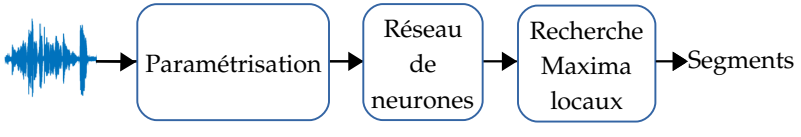


FIGURE 1 – Schéma du système de segmentation en phones

2.1 Paramétrisation

Suite à différents essais de paramètres temporels et fréquentiels, nous avons opté pour des bancs de filtres, calculés sur le signal découpé en fenêtres de 16 ms, avec un pas de 4 ms. Nous extrayons les coefficients FBANK, que nous donnons en entrée du réseau de neurones. Il est rappelé que le processus d'extraction des FBANK est fondé sur la transformation de l'amplitude spectrale grâce à un banc de filtres. Celui-ci est caractérisé par des filtres triangulaires, répartis de manière linéaire selon l'échelle Mel. Le logarithme des énergies obtenu après filtrage est calculé pour obtenir les FBANK.

2.2 Réseaux de neurones

Les CNN sont très efficaces en reconnaissance de formes : plus de 99% de reconnaissance correcte sur des chiffres manuscrits (MNIST) (LeCun *et al.*, 1998), par exemple. Le MLP peut parvenir à des résultats similaires avec davantage de couches : 12 couches totalement connectées contre 6 (1 couche de convolution et 5 totalement connectées) pour un CNN (Golik *et al.*, 2015). Dans cet article, nous comparons ces deux types de réseaux de neurones (CNN et MLP), avec comme objectif de reconnaître les variations dans le spectrogramme marquant un changement de phones, changement qui varie selon les phones considérés.

Le réseau de neurones voit la tâche de segmentation comme une tâche de classification binaire : présence / absence de frontière. Classiquement, lors de l'attribution d'une classe à un individu, il calcule d'abord pour chaque classe la probabilité que l'individu étudié lui appartienne, puis il indique en sortie la classe la plus probable. Cependant, cette dernière étape rencontre deux difficultés : les deux classes de frontière (présence, absence) étant réparties en des proportions inégales (i.e. 1/5, 4/5), les probabilités en sortie sont plus difficilement favorables à la présence de frontière. De plus, lorsqu'une fenêtre a une probabilité élevée d'être une frontière, alors ses voisines ont de grandes chances de l'être elles aussi. Pour éviter cela, il faudrait que le réseau considère les probabilités des fenêtres voisines avant de prendre une décision, ce qui est davantage envisageable avec un réseau de type récurrent. Nous avons choisi de traiter nous-même les probabilités en sortie du réseau pour en déduire les bornes des segments des phones à l'aide d'une méthode de recherche de maxima locaux.

2.3 Recherche de maxima locaux

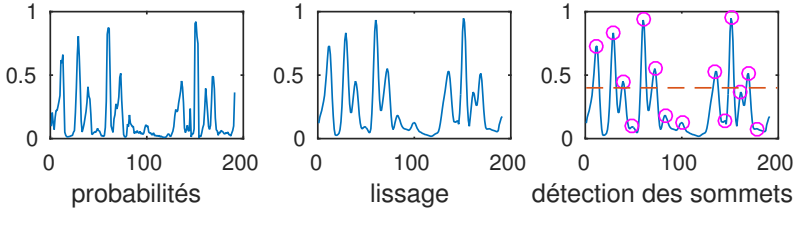


FIGURE 2 – Illustration de notre recherche de maxima locaux sur un enregistrement constitué de 200 fenêtres d'analyse

La figure 2 illustre le processus de recherche des maxima locaux. Pour chaque fenêtre d'analyse, le

réseau de neurones calcule une probabilité que celle-ci contienne une frontière (passage d'un phone à un autre). Chaque enregistrement donne lieu à une courbe de probabilités (200 valeurs sur la figure 2). Pour éviter de détecter des variations locales dues au bruit, nous lissons la courbe à l'aide d'une convolution avec une fenêtre de Hamming de petite taille (5, dans notre cas). Nous détectons ensuite les sommets (maxima locaux) et nous ne conservons que ceux supérieurs à un seuil. La valeur du seuil peut varier selon les besoins de privilégier la précision, le rappel ou la F-mesure. Le seuil maximisant la F-mesure correspond environ au seuil sélectionnant approximativement 12 phones par secondes pour le corpus de parole conversationnelle BUCKEYE et 9 pour le corpus lu de Xitsonga.

3 Corpora et métriques d'évaluation

Nous avons utilisé le corpus d'anglais américain appelé BUCKEYE (Pitt *et al.*, 2007), constitué de parole spontanée (enregistrements de radio) d'une quarantaine de locuteurs différents (hommes, femmes, jeunes, personnes âgées) avec environ 30 minutes de temps de parole pour chacun. Ce corpus est décrit en détails dans (Kiesling *et al.*, 2006). La qualité des annotations manuelles a été évaluée par les créateurs du corpus. Un accord inter-annotateurs a été calculé : il est de l'ordre de 62% de F-mesure pour la segmentation avec 10 ms de marge (décalage). Le pourcentage monte à 79% pour un décalage de 20 ms (Raymond *et al.*, 2002). La durée médiane des phonèmes est d'environ 70 ms, avec une soixantaine de phonèmes différents annotés, chiffre supérieur à la quarantaine habituellement référencée en anglais notamment à cause de prononciations particulières que les auteurs de BUCKEYE ont choisi d'isoler dans des classes différentes, pour les nasales en particulier. En nous basant sur le découpage du challenge *Zero Resource Speech*, nous avons divisé le sous-corpus d'apprentissage en deux parties : un sous-corpus d'entraînement (BUCKEYE-TRAIN, 75%, 10 heures, 20 locuteurs), un corpus de développement (BUCKEYE-DEV, 25%, 3 heures, 6 locuteurs), et nous avons conservé la partie officielle de test (BUCKEYE-TEST, 5 heures, 12 locuteurs) telle quelle.

Le corpus en langue Xitsonga (van Heerden *et al.*, 2013) est composé de courtes phrases lues, enregistrées sur Smartphone hors studio. Nous avons utilisé près de 500 phrases, avec en tout 10000 exemples de phonèmes annotés manuellement, issus de la base de données du même challenge *Zero Resource Speech*. La durée médiane des phones est d'environ 90 ms et il y a 49 phones différents.

Une certaine marge d'erreur est tolérée lors de l'attribution de la frontière. Nous avons utilisée deux marges différentes : la marge la plus courante dans la littérature (20 ms) et une marge plus petite de 10 ms parfois aussi trouvée. Pour évaluer les résultats, nous avons utilisé les métriques classiques de précision, rappel et F-mesure. Selon le seuil choisi pour la recherche des maxima locaux, nous repérons plus ou moins de frontières et obtenons des scores différents. Les courbes DET (*Detection Error Trade-off*), ayant en abscisse le taux de faux positifs et en ordonnée le taux de faux négatifs, nous permettront de visualiser les différents résultats selon les seuils testés (Martin *et al.*, 1997).

4 Expériences

4.1 Comparaison de différentes configurations sur BUCKEYE-DEV

Dans le cadre de cet article, nous avons utilisé Theano (Bastien *et al.*, 2012) et Lasagne (Dieleman *et al.*, 2015) pour la mise en œuvre des modèles. Avant de chercher à optimiser les paramètres, nous

avons tout d'abord comparé le CNN et le MLP. En utilisant les bancs de filtres en entrée, le CNN s'est montré pertinent (taux d'apprentissage=0.007, coefficient de régularisation=0.9). Le MLP a eu, quant à lui, besoin de la dérivée des bancs de filtres pour obtenir des résultats intéressants avec un modèle peu profond. Cette dérivée n'a pas été pertinente pour le CNN et une étude a montré que de sa première couche de convolution effectuait elle-même plusieurs approximations d'une dérivée temporelle. Nous avons donc optimisés les paramètres pour chacun des deux réseaux (nombre de couches, de neurones, dimension des filtres de convolution) avant de faire le choix le plus pertinent.

Suite au choix d'ajouter la dérivée des bancs de filtres en entrée du MLP, celui-ci s'est montré peu sensible à son nombre de couches et nous avons restreint ce nombre à 3 couches cachées. Cependant, le nombre de neurones a eu beaucoup plus d'influence, avec un maximum de performance autour de 300 neurones (cf. figure 3), mais pour une augmentation de seulement 1% de la F-mesure, par rapport à un modèle uniquement constitué de 50 neurones. Le CNN est optimal, quant à lui, entre 50 et 400 neurones pour les couches totalement connectées. Le nombre de filtres de ses couches de convolution a un impact de l'ordre de 1% à 2%, en absolu. Passer de 15 à 120 filtres permet ainsi de gagner 1,2%.

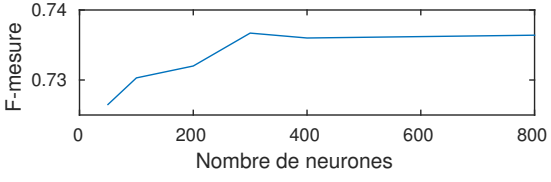


FIGURE 3 – Evolution de la F-mesure en fonction du nombre de neurones des couches cachées du MLP sur BUCKEYE-DEV

Le nombre de fenêtres voisines considérées s'est avéré être l'un des paramètres les plus importants : les changements de phones se repèrent notamment grâce au contexte. Le MLP supporte moins bien l'augmentation du nombre de données (allant avec l'augmentation du nombre de voisins) que le CNN, dont les couches de convolution effectuent visiblement un premier traitement pertinent et réducteur. La figure 4 illustre l'importance de la taille du contexte pour la tâche de segmentation : les résultats s'améliorent de manière visible avec l'augmentation du nombre de voisins. Nous avons choisi 18 voisins (84 ms), l'augmentation au-delà étant très faible par rapport à l'augmentation de la complexité.

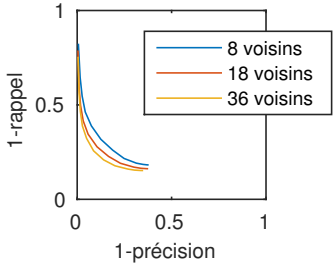


FIGURE 4 – Courbe DET avec un CNN en fonction du voisinage considéré sur BUCKEYE-DEV

La taille du voisinage étant un paramètre influent sur le réseau, nous comparons nos deux modèles (CNN et MLP) en fonction de son évolution. Sur la figure 5, nous voyons que le CNN s'avère plus efficace que le MLP : nous l'utiliserons donc uniquement celui-ci dans la suite du document.

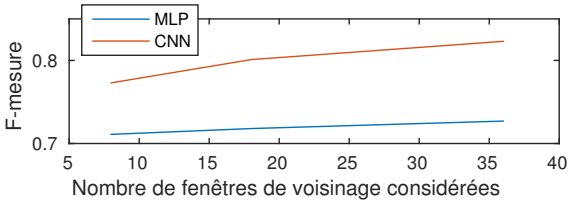


FIGURE 5 – Comparaison des F-mesures du MLP et du CNN selon le nombre de fenêtres de voisinage sur BUCKEYE-DEV

4.2 Résultats sur BUCKEYE-TEST

Avec un CNN composé de 2 couches de convolution dotées de 60 filtres et d'une couche totalement connectée de 200 neurones, nous avons obtenu une F-mesure de 68% pour une tolérance de 10 ms. Nous pouvons obtenir une précision proche de 90%, si nous acceptons de ne trouver qu'un tiers des frontières, ou bien un rappel de 72% avec la moitié des détections erronées (cf. table 1).

Précision	Rappel	F-mesure
0.52	**0.72**	0.61
0.71	0.65	**0.68**
0.94	0.16	0.27

TABLE 1 – Résultats sur BUCKEYE-TEST pour 4 valeurs de seuil et 10 ms de tolérance

La figure 6 est un exemple de résultat obtenu par le réseau de neurones, montrant la courbe des probabilités superposée au spectrogramme du signal. Les valeurs élevées de la courbe correspondent effectivement à des variations dans le spectre et sont corrélées avec les frontières.

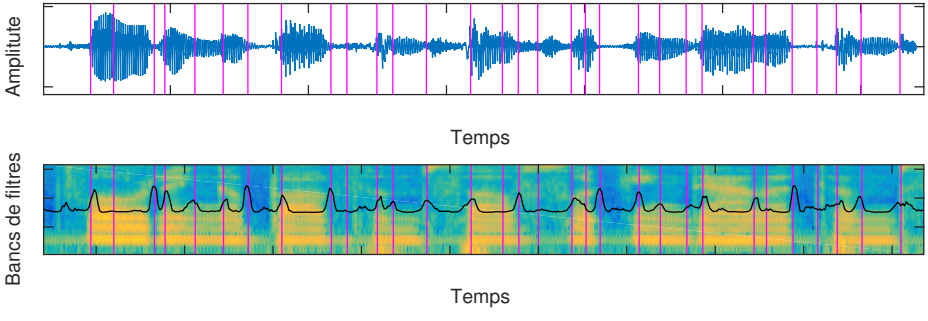


FIGURE 6 – Probabilités en sortie du CNN pour la segmentation sur BUCKEYE-TEST

Nous avons analysé les taux de détection des frontières de quelques phones parmi les plus fréquents (cf. table 2). Nous constatons que les frontières des phones courts avec une forte attaque tels que /g/ ou /k/ sont souvent trouvées alors que le réseau rencontre davantage de difficultés pour le /l/ et le /r/. Ainsi par exemple pour 10 ms de marge, la frontière entre /ao/ et /l/ n'est trouvées que dans 9% des

cas, 15% pour celle entre /aw/ et /r/. Les frontières entre deux voyelles ont aussi de mauvais scores : 8% entre /ow/ et /ay/, 11% entre /er/ et /ay/.

	/r/	/l/	/ao/	/uh/	/g/
% débuts segments détectés	46	45	73	62	81
% fins segments détectées	49	51	39	76	81

TABLE 2 – Analyse des résultats pour 5 phones différents – BUCKEYE-TEST, 10 ms de tolérance

Les résultats en segmentation automatique sont proches de l'erreur constatée entre les annotateurs humains. La table 3 montre même que notre système est plus précis lorsqu'il localise une frontière juste : nous avons une meilleure F-mesure pour 10 ms de tolérance d'erreur et son augmentation entre 10 ms et 20 ms est plus faible que pour celle observée entre les annotateurs.

	annotateurs	CNN
10 ms	0.62	0.68
20 ms	0.79	0.79

TABLE 3 – Comparaison de F-mesures entre l'accord inter-annotateurs et le CNN – BUCKEYE-TEST

4.3 Segmentation avec peu de données d'apprentissage

La segmentation ne séparant les données qu'en deux classes différentes, nous pouvons espérer que peu d'échantillons suffisent pour l'apprentissage, et qu'utiliser un modèle appris sur une langue avec beaucoup de données peut quand même détecter des frontières si nous l'utilisons pour une autre langue, pour laquelle peu de données sont disponibles.

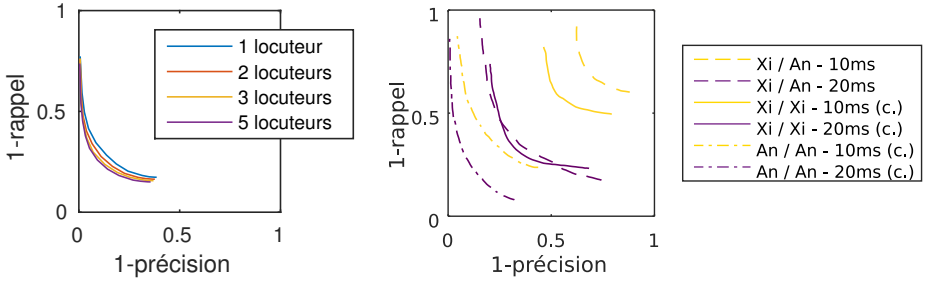


FIGURE 7 – Courbes DET sur des cas difficiles : **à gauche** : peu de locuteurs, **à droite** : apprentissage sur une langue mieux dotée. Notations : (c.) résultats obtenus en validation croisée sur peu de données selon la marge de tolérance (ms), Xi pour Xitsonga, An pour Anglais, [corpus de test]/[corpus d'apprentissage]

La figure 7 montre la bonne adaptation du CNN à des cas peu dotés. Sur le graphique de gauche, nous remarquons qu'apprendre sur un unique locuteur donne bien évidemment des résultats légèrement inférieurs à ceux obtenus à l'aide de plusieurs locuteurs. En utilisant les données de 3 ou 5 locuteurs, nous avons obtenu des résultats très proches et l'amélioration au-delà de 5 locuteurs est presque inexistante.

Sur le graphique de droite, se trouvent différents résultats du CNN dans le cas du Xitsonga, langue peu dotée sur laquelle nous avons effectuée les tests. Les résultats sur le corpus de Xitsonga sont moins bons que sur le corpus d'anglais dans des conditions d'apprentissage similaires (en validation croisée sur le même nombre d'échantillons, de 4 locuteurs différents). Ceci peut s'expliquer par les conditions d'enregistrement différentes : les enregistrements de Xitsonga ayant été réalisés avec des smartphones, hors studio. Un résultat intéressant qui apparaît est que le modèle appris sur le petit corpus de Xitsonga donne des résultats meilleurs avec une tolérance de 10 ms d'erreur que le modèle appris sur BUCKEYE-TRAIN, mais similaires pour une tolérance de 20 ms. Nous pouvons donc en conclure que le modèle appris sur le grand corpus d'anglais se montre surtout moins précis lors de l'attribution des frontières des phones du Xitsonga.

Afin de mieux appréhender les résultats, nous avons affiché un exemple sur la figure 8.

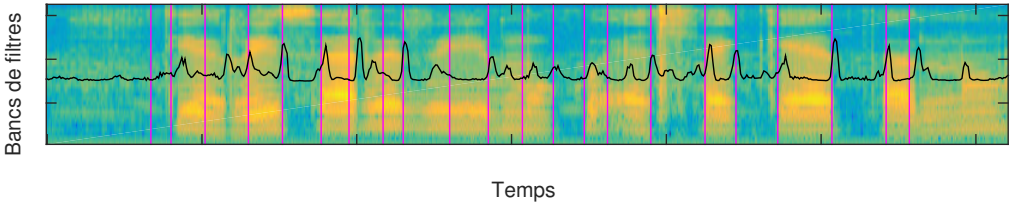


FIGURE 8 – Probabilités de frontières en sortie du CNN appris sur BUCKEYE-TRAIN pour le corpus Xitsonga

5 Conclusion

Dans cet article, nous avons décrit des résultats expérimentaux de segmentation automatique de la parole en phones à l'aide de différents réseaux de neurones (CNN et MLP). Sur les enregistrements d'anglais américain issu du corpus BUCKEYE, nos réseaux de neurones ont obtenu des résultats assez remarquables : 68% de F-mesure pour notre meilleur système de segmentation automatique contre 62% pour l'accord inter-annotateurs, avec une tolérance de 10 ms sur la localisation des frontières de phones. De plus, les modèles ont fait preuve d'une bonne adaptation à des cas particuliers difficiles : peu de données d'apprentissage et application à une langue différente de celle de l'apprentissage. En particulier, des performances similaires ont été obtenues sur un petit corpus de la langue peu dotée Xitsonga en utilisant : 1) un modèle entraîné sur l'anglais américain, 2) un modèle entraîné sur un sous-corpus de petite taille de la langue peu dotée. Ce résultat nous fait supposer que des modèles entraînés sur des langues disposant de grandes quantités de données peuvent être utilisés avec des langues peu dotées en première approche (Renshaw *et al.*, 2015). Pour la segmentation de langues peu dotées, telles que le Xitsonga, nous envisageons de réaliser des expériences d'apprentissage semi-supervisé comme le *bootstrap*, en utilisant des modèles entraînés sur l'anglais américain et d'autres grands corpora.

Références

BASTIEN F., LAMBLIN P., PASCANU R., BERGSTRA J., GOODFELLOW I. J., BERGERON A., BOUCHARD N. & BENGIO Y. (2012). Theano : new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop.

BHARGAVA M. & ROSE R. (2015). Architectures for deep neural network based acoustic models defined over windowed speech waveforms. p. 6–10.

DIELEMAN S., SCHLÜTER J., RAFFEL C., OLSON E., SØNDERBY S. K., NOURI D., MATURANA D., THOMA M., BATTENBERG E., KELLY J., FAUW J. D., HEILMAN M., DIOGO149, MCFEE B., WEIDEMAN H., TAKACSG84, PETERDERIVAZ, JON, INSTAGIBBS, RASUL D. K., CONGLIU, BRITEFURY & DEGRAVE J. (2015). Lasagne : First release.

GOLIK P., TÜSKE Z., SCHLÜTER R. & NEY H. (2015). Convolutional neural networks for acoustic modeling of raw time signal in lvcsr. p. 26–30.

JANSEN A., DUPOUX E., GOLDWATER S., JOHNSON M., KHUDANPUR S., CHURCH K., FELDMAN N., HERMANSKY H., METZE F., ROSE R. *et al.* (2013). A summary of the 2012 jhu clsp workshop on zero resource speech technologies and models of early language acquisition.

JOSHI S., DEO N. & RAO P. (2015). Vowel mispronunciation detection using dnn acoustic models with cross-lingual training. p. 697–701.

KIESLING S., DILLEY L. & RAYMOND W. D. (2006). The variation in conversation (vic) project : Creation of the buckeye corpus of conversational speech. p. 55–97.

LECUN Y., BOTTOU L., BENGIO Y. & HAFFNER P. (1998). Gradient-based learning applied to document recognition. p. 2278–2324.

LEE C.-Y., O'DONNELL T. J. & GLASS J. (2015). Unsupervised lexicon discovery from acoustic input. *Transactions of the Association for Computational Linguistics*, **3**, 389–403.

MARTIN A., DODDINGTON G., KAMM T., ORDOWSKI M. & PRZYBOCKI M. (1997). *The DET curve in assessment of detection task performance*. Rapport interne, DTIC Document.

PITT M., DILLEY L., JOHNSON K., KIESLING S., RAYMOND W., HUME E. & FOSLER-LUSSIER E. (2007). Buckeye corpus of conversational speech (2nd release). www.buckeyecorpus.osu.edu.

RAYMOND W. D., PITT M., JOHNSON K., HUME E., MAKASHAY M., DAUTRICOURT R. & HILTS C. (2002). An analysis of transcription consistency in spontaneous speech from the buckeye corpus.

RENSHAW D., KAMPER H., JANSEN A. & GOLDWATER S. (2015). A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge. p. 3199–6303.

VAN HEERDEN C., DAVEL M. & BARNARD E. (2013). The semi-automated creation of stratified speech corpora.

VERSTEEGH M., THIOLLIÈRE R., SCHATZ T., CAO X. N., ANGUERA X., JANSEN A. & DUPOUX E. (2015). The zero resource speech challenge 2015. p. 3169–3173.

L'invasivité phonologique dans le traitement des anglicismes : une étude quantitative de trois langues

Tomáš Duběda
Institut de la langue tchèque, République tchèque
dubeda@ff.cuni.cz

RÉSUMÉ

Dans la présente étude, nous analysons, dans une perspective typologique, l'adaptation phonologique des anglicismes dans trois langues (français, allemand et tchèque). La classification des formes phonologiques, qui s'appuie sur un système de huit principes d'adaptation, a pour but d'établir le degré d'« invasivité phonologique » propre à chaque langue. L'approximation phonologique (substitution de phonèmes natifs aux phonèmes étrangers) semble être le principe fondamental dans les trois langues analysées, alors que la prononciation orthographique (phonétisation des graphèmes) intervient avant tout en français. La prononciation authentique (imitation phonologique de la langue source) n'est active qu'en allemand. Les mécanismes d'approximation phonologique sont plus invasifs en français que dans les deux autres langues, et ce notamment en ce qui concerne le système vocalique. Globalement, l'invasivité phonologique semble augmenter dans l'ordre allemand – tchèque – français.

ABSTRACT

Phonological invasiveness in the treatment of loanwords. A quantitative study of three languages.
In the present paper, we analyse the phonological adaptation of Anglicisms in three languages (French, German and Czech) from a typological perspective. The classification of phonological forms, based on a system of eight adaptation principles, aims at quantifying the degree of "phonological invasiveness" for each of the languages. Phonological approximation (substitution of foreign phonemes with native ones) seems to be the fundamental principle in all three languages analysed, while spelling pronunciation (phonetisation of graphemes) is observed especially in French. Authentic pronunciation (phonological imitation of the source language) is only active in German. The mechanisms of phonological approximation are more invasive in French than in the other two languages, particularly with regard to the vowel systems. Globally, the phonological invasiveness increases in the order German – Czech – French.

MOTS-CLÉS : Phonologie, adaptation phonologique, emprunts, anglicismes, français, allemand, tchèque.

KEYWORDS: Phonology, phonological adaptation, loanwords, Anglicisms, French, German, Czech.

1 Introduction

Toutes les langues européennes connaissent le mécanisme de l'emprunt, consistant à adopter dans leur lexique des mots venus d'autres langues. Si certaines d'entre elles, comme le finnois, contiennent remarquablement peu de mots d'origine étrangère (Genzor, 2015), d'autres ont subi dans leur histoire une forte hybridisation : ainsi, pour l'anglais, on estime le nombre d'emprunts à 70 %, quoique la partie centrale du vocabulaire reste anglo-saxonne (Hogg, Denison, 2008). Il est quelque peu paradoxal que l'anglais, après s'être enrichi d'une telle quantité d'emprunts, est devenu au cours du XX^{e} siècle, grâce à la situation géopolitique, « de loin le plus grand exportateur lexical » (Görlach, 2001) : le *Dictionary of European Anglicisms* (op. cit.) nous donne une idée de l'ampleur du phénomène.

Le présent texte se penche sur un des aspects formels de l'adoption des anglicismes, à savoir leur adaptation phonologique. Lors du passage de la langue donneuse vers la langue emprunteuse, les mots subissent, en règle générale, une transformation phonologique qui facilite leur fonctionnement dans cette dernière langue. Cette « réparation phonologique » (Calabrese et Wetzels, 2009), qui a pour but de résoudre les tensions entre les deux systèmes phonologiques, peut être décrite à travers les huit principes suivants (Duběda et al., 2014) :

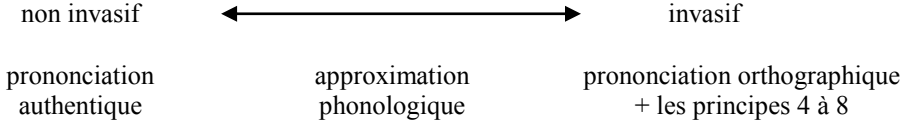
1. **Approximation phonologique** (substitution aux phonèmes originaux des phonèmes les plus proches de la langue cible, normalisation phonotactique et prosodique) : angl. *fan club* [ˈfænklʌb] > fr. *fan-club* [fanˈklœb].
2. **Prononciation orthographique** (application des règles de conversion graphie-phonie de la langue cible) : angl. *laser* [ˈleɪzə] > fr. *laser* [laˈzɛːʁ] (et non pas *[leˈzːœʁ], comme cela aurait été le cas si le principe d'approximation avait était appliqué).
3. **Prononciation partiellement ou totalement authentique du mot**, qui est utilisée surtout pour les noms propres et plus généralement pour les emprunts moins intégrés, et qui introduit dans la langue emprunteuse des éléments phonétiques étrangers.
4. **Analogie avec la langue source :** par exemple, le mot anglais *sweatshirt* [ˈswetʃɜːt] est parfois prononcé en français comme [switˈʃœʁt], par analogie avec des mots où le digraphe *ea* se prononce comme [iː] (*read, meat, deal*), et peut-être aussi par analogie avec le mot *sweet*.
5. **Analogie avec la langue cible :** par exemple, le mot *aborigène* est parfois prononcé comme [aʁbɔʁiˈʒɛn], sous l'influence du mot *arbre*, jugé comme appartenant au même champ sémantique.
6. **Prononciation influencée par une troisième langue :** par exemple, *Eiffelova věž* « Tour Eiffel » se prononce en tchèque comme [ˈʔajfɛlova ˈvjɛʃ], sous l'influence des règles de prononciation allemande (*ei* [aɪ]).
7. **Prononciation influencée par les universaux :** par exemple, l'adaptation prédominante du mot français *tartelette* en tchèque est *tartaletka* [ˈtartalɛtka], où la présence d'un [a] dans la deuxième syllabe est explicable par le principe d'harmonie vocalique.
8. **Prononciation sans motivation apparente :** par exemple, le mot tchèque *brožura* [ˈbrɔʒura] « brochure » contient un [ʒ] à la place d'un [ʃ], alors que le [ʃ] est un phonème courant en tchèque et que le voisement intervocalique est un principe très marginal en tchèque ; le passage par l'allemand ne permet pas non plus d'expliquer ce phénomène (*Broschüre* [bʁɔˈʃyːʁə]).

Deux ou plusieurs principes peuvent agir simultanément au sein du même mot : dans la prononciation française du mot *boy-scout* [bɔjˈskut], la première diphtongue est adaptée selon le

principe 1, et la deuxième selon le principe 2. De plus, différents principes peuvent engendrer des prononciations concurrentes : *pipeline* [pipˈlin] vs. [pajpˈlajn] (principes 2 vs. 1).

Les langues diffèrent entre elles quant aux principes d'adaptation qu'elles préfèrent. Si l'on considère l'adaptation des anglicismes dans les trois langues qui font l'objet de notre étude, on constate en premier lieu que le français est assez réticent à la prononciation authentique, sauf le phonème /ŋ/ (*bowling, smoking*), qu'il a adopté dans son système phonologique. Dans les variantes régionales du français, notamment en Suisse et en Belgique, cette résistance est peut-être moindre. Le français fait donc surtout appel à l'approximation phonologique (Retman, 1978) : *feed-back, lady*, avec de nombreux cas de prononciation orthographique : *label, scout*. L'allemand, de son côté, est sensiblement plus ouvert aux prononciations authentiques, comme en témoignent les ouvrages linguistiques (Lange, 2015), mais aussi la pratique des médias, notamment en ce qui concerne la prononciation des noms propres d'origine étrangère. Le dictionnaire *Duden – Universalwörterbuch* inclut dans la liste des symboles de transcription plusieurs sons d'origine française [ã, ãː, ɛ, ɛ̃, o, õ, œ, œ̃, ʒ] et réserve une section spéciale aux sons faisant partie de la « prononciation anglaise » : [ɑː, æ, ʌ, ð, θ, w]. Malgré cela, le principe d'approximation phonologique semble déterminant aussi pour l'allemand : *happy end* [ˈhɛpiˈʔɛnt], *Mixer* [ˈmɪksɐ]. Le tchèque est décrit comme une langue favorisant l'approximation phonologique (Romportl, 1978 ; Duběda et al., 2014) : *feedback* [ˈfiːdbɛk], *copyright* [ˈkɔpɪrajt]. Malgré cela, cette langue a incorporé dans son système phonologique plusieurs éléments d'origine étrangère, pour la plupart gréco-latine : /f/, /g/, /ɔː/, /au/, /ɛu/, mais aussi anglaise : /ʤ/. La forte régularité de l'orthographe tchèque semble influer sur les mots d'origine étrangère en ce sens que les emprunts bien intégrés ont tendance à s'adapter orthographiquement : *svetr* [ˈsvɛtr̩], *tramvaj* [ˈtramvaj].

Le degré d'adaptation qui caractérise chaque langue peut être exprimé en termes d'« invasivité phonologique ». Celle-ci est déterminée, en premier lieu, par les principes d'adaptation, selon l'échelle suivante :



En second lieu, cette invasivité dépend de la nature de l'approximation phonologique, c'est-à-dire de la distance entre les phonèmes originaux et les phonèmes qui leurs sont substitués. Ainsi, l'adaptation allemande du mot anglais *dispatcher* [dɪsˈpæʧə] mène à une forme qui reste assez proche de l'original : [ˈdɪspɛʧɐ], alors que l'adaptation française s'éloigne davantage de la structure phonologique de départ : [dispaˈʧːœʁ], quoiqu'il s'agisse dans les deux cas de l'application des règles habituelles d'approximation phonologique. Ce type d'invasivité est donc fonction de la distance qui sépare le système phonologique de la langue donneuse de celui de la langue emprunteuse.

2 Objectif et hypothèses

L'objectif du présent article est de quantifier le degré d'invasivité phonologique, telle qu'elle a été définie plus haut, dans le traitement des anglicismes dans trois langues européennes appartenant à des groupes différents : le français (galloroman), l'allemand (germanique occidental) et le tchèque

(slave occidental). Nos hypothèses, toutes exprimant divers aspects de l'invasivité phonologique, sont les suivantes :

1. Le principe d'approximation phonologique prédomine dans les trois langues.
2. Le français favorise la prononciation orthographique plus que les deux autres langues.
3. L'allemand favorise la prononciation authentique plus que les deux autres langues.
4. L'approximation phonologique est moins invasive en allemand que dans les deux autres langues, du fait de la ressemblance des deux systèmes phonologiques.

3 Matériaux

L'étude quantitative est basée sur un échantillon d'anglicismes tiré du *Dictionary of European Anglicisms* (Görlach, 2001). Dans un premier temps, nous avons inclus dans la liste les 852 entrées qui sont traitées dans le dictionnaire de manière approfondie, et qui sont facilement repérables grâce à leur organisation graphique. Ces « entrée phares » correspondent à des anglicismes fréquents et largement attestés, donc particulièrement pertinents pour notre étude. Dans un deuxième temps, nous avons vérifié pour chaque entrée si elle est présente dans les dictionnaires suivants :

– Français : *Le Petit Robert de la langue française*
– Allemand : *Duden. Universalwörterbuch*
– Tchèque : *Nový akademický slovník cizích slov* [Nouveau dictionnaire académique des emprunts]. Comme le dernier grand dictionnaire généraliste du tchèque date de 1971, nous avons opté pour un dictionnaire spécialisé, qui reflète mieux l'état actuel de la langue.

Nous avons retenu pour notre analyse les 219 entrées qui étaient présentes dans les trois dictionnaires. Cet échantillon, dans sa version quadrilingue (anglais – français – allemand – tchèque), complété d'informations sur la prononciation, nous offre une base empirique strictement comparable (voir Table 1 pour un exemple). En ce qui concerne la prononciation anglaise, nous nous sommes référé au *Cambridge English Pronouncing Dictionary* (prononciation britannique). Un grand nombre d'entrées tchèques et certaines entrées allemandes à prononciation régulière n'étaient pas pourvues de transcription phonétique dans les dictionnaires respectifs, ou étaient pourvues d'une transcription partielle ; ces entrées ont été transcrites par l'auteur.

Anglais		Français		Allemand		Tchèque	
jackpot	ˈʤækpɒt	*jackpot*	ʤakˈpɔt ʒakˈpɔt	*Jackpot*	ˈʤækpɔt	*jackpot*	ˈʤɛkpɔt
jamboree	ˌʤæmbəˈriː	*jamboree*	ʒɑ̃bɔˈʁe, ʒambɔˈʁi	*Jamboree*	ʤæmbəˈriː	*jamboree*	ˈʤɛmbɔriː
jet	ˈʤet	*jet*	ˈʤɛt	*Jet*	ˈʤɛt	*jet*	ˈʤɛt
jogging	ˈʤɒgɪŋ	*jogging*	ʤɔˈgiŋ	*Jogging*	ˈʤɔgɪŋ	*jogging*	ˈʤɔgɪŋk
jukebox	ˈʤuːkbɒks	*juke-box,* *jukebox*	ʒykˈbɔks, ʤukˈbɔks	*Jukebox*	ˈʤuːkbɔks	*juke-box*	ˈʤuːgbɔks

TABLE 1 : Extrait de l'échantillon (lettre *J*). La transcription du tchèque a été convertie en API. Les symboles d'accent ont été ajoutés là où ils n'étaient pas indiqués.

Chaque forme phonologique adaptée a été ensuite classifiée selon le ou les principes d'adaptation qui en sont responsables. Les six principales catégories attribuées sont les suivantes :

– APPROX – approximation phonologique : fr. *feed-back* [fidˈbak]
– ORTH – prononciation orthographique : fr. *label* [laˈbɛl]
– AUTH – prononciation authentique : all. *Hardware* [ˈhɑːdweə]
– APPROX=ORTH – forme phonologique expliquable à la fois par l'approximation phonologique et par le principe orthographique : tch. *drift* [ˈdrɪft]
– APPROX+ORTH – combinaison des deux principes dans la même entrée : *boy-scout* [bɔjˈskut]
– APPROX/ORTH – deux variantes coexistantes : fr. *pipeline* [pipˈlin / pajpˈlajn]
– APPROX/AUTH – deux variantes coexistantes : all. *Badminton* [ˈbɛtmɪntən / ˈbædmɪntən]

À part cela, nous avons identifié, en allemand uniquement, quatre autres catégories qui réunissent deux ou trois principes et qui correspondent à un très petit nombre d'observations, par exemple *Steak* [ˈsteːk / ˈʃteːk] APPROX/APPROX+ORTH (deux variantes coexistantes, l'une basée sur l'approximation phonologique et l'autre sur une combinaison de l'approximation phonologique et de la prononciation orthographique). À cela s'ajoute la catégorie ANOM (4 observations en tout), qui correspond aux prononciations ne rentrant dans aucune des catégories susmentionnées et souvent influencées par des analogies, par exemple fr. *steward* [stiˈwaʁt]. Toutes ces catégories, dont l'effectif total est de 13, seront réunies ci-après sous l'étiquette « AUTRES ».

Notons en marge que les prononciations authentiques indiquées dans le *Duden* correspondent aux variantes britanniques, avec un choix de symboles qui diffère légèrement de la norme. Les différences par rapport au *Cambridge English Pronouncing Dictionary* sont les suivantes : *CEPD* [i ɒ eə əʊ] ; *Duden* [ɪ ɔ ɛə oʊ]. Ces détails n'ont aucune importance pour notre étude, qui est phonologique plutôt que phonétique.

En basant notre analyse sur les prononciations indiquées dans les ouvrages lexicographiques, nous avons fait un choix méthodologique qui nécessite un commentaire : on sait que les mots d'origine étrangère montrent, en règle générale, une variabilité phonologique qui est plus importante que celle des mots natifs (Retman, 1978 ; Muhvić-Dimanovski, 1995 ; Duběda et al., 2014) ; l'analyse se limite donc aux formes considérées comme correctes et recommandables, et adoptées comme telles par des dictionnaires qui font autorité. Nous sommes cependant d'avis que ces formes reflètent fidèlement le type de traitement phonologique qui est typique de chacune des langues étudiées.

4 Analyse

4.1 Les principes d'adaptation

La Figure 1 montre la répartition des entrées en fonction des principes d'adaptation observés.

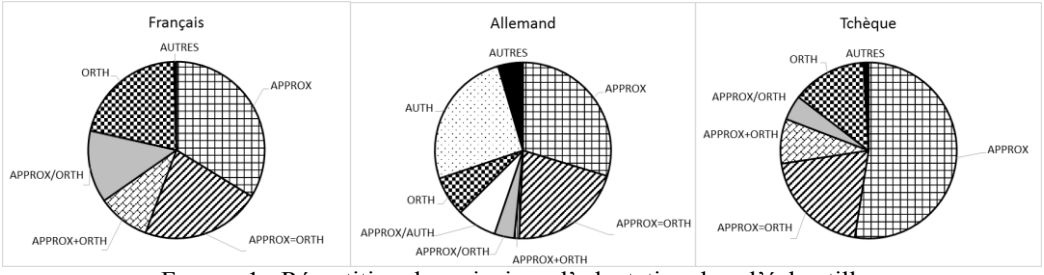


FIGURE 1 : Répartition des principes d'adaptation dans l'échantillon

Les données confirment la prééminence du principe d'approximation phonologique (APPROX) dans les trois langues analysées, corroborant ainsi l'hypothèse 1. Ce principe est responsable à lui seul d'environ un tiers des anglicismes adaptés en français et en allemand (34 % et 30 % respectivement), et de plus d'une moitié en tchèque (53 %). Quant à la prononciation purement orthographique (ORTH), sa fréquence augmente dans l'ordre allemand (7 %) < tchèque (14 %) < français (21 %). L'hypothèse 2 peut donc également être confirmée.

À part l'approximation phonologique et la prononciation orthographique dans leur forme pure, l'échantillon inclut un nombre important de mots dont l'adaptation est basée sur la synergie entre ces deux principes. Tout d'abord, entre 20 % et 22 % des entrées correspondent à la catégorie APPROX=ORTH : l'adaptation de ces mots, grâce à leur structure phonologique « passe-partout » (*drift, flop, kit*…), aboutit au même résultat, qu'on envisage l'approximation phonologique ou le principe orthographique. Ensuite, la catégorie APPROX+ORTH (adaptation « hybride »), inclut environ un dixième de l'échantillon en français et en tchèque, mais est presque absente de l'allemand. Finalement, la catégorie APPROX/ORTH (adaptation parallèle) est sensiblement plus importante en français que dans les deux autres langues.

La prononciation authentique (AUTH) est absente des dictionnaires français et tchèque, alors qu'elle est recommandée par le dictionnaire allemand comme seule variante pour 24 % des entrées étudiées, et comme une variante alternative (APPROX/AUTH) pour 7 % de ces entrées. L'hypothèse 3 s'en trouve donc confirmée.

4.2 L'approximation phonologique

L'approximation phonologique est un procédé consistant à « greffer » le système phonologique de la langue donneuse sur le système phonologique de la langue emprunteuse en fonction de leur catégorisation perceptive. Les règles d'approximation des monophtongues identifiées dans notre échantillon sont résumées dans la Table 2.

Anglais				Français			Allemand				Tchèque			
iː			uː	i		u	iː			uː	iː			uː
ɪ			ʊ				ɪ			ʊ	ɪ			u
e	ə	ɜː	ɔː	e/ɛ	œ	o/ɔ	ɛ	ɐ/ə	øː/œ	ɔ(ː)	ɛ	ɛ/r̩	ɛː	ɔː
	ʌ		ɒ			ɔ		a		ɔ		a		ɔ
æ		ɑː		a				a(ː)				aː		

TABLE 2 : Approximation des monophtongues dans l'échantillon. Pour le français, le choix des variantes [e/ɛ], [o/ɔ] est contextuel. Le choix [ɐ/ə] pour l'allemand et le choix [ɛ/r̩] pour le tchèque dépendent de la rhoticité sous-jacente. L'approximation du [ɜː] en allemand n'est attestée qu'une seule fois (*Girl*), et l'approximation du [ɔː], deux fois (*Heliport, Softball*).

Le système anglais à 12 monophtongues se voit refondu en un système qui comporte 7 éléments en français, 11 en allemand et 11 en tchèque. La forte réduction du nombre de phonèmes en français découle avant tout du fait que cette langue est incapable, à la différence des deux autres, d'exprimer la longueur phonologique. Les voyelles longues fusionnent alors avec les voyelles brèves les plus proches. Les voyelles centrales [ə ɜː ʌ] subissent également une fusion importante. Les systèmes allemand et tchèque comportent une seule fusion totale : [e æ], et plusieurs fusions partielles : [ɔ ɔː], [a aː], [ɛ ɛ/ɽ]. Il est évident que l'approximation phonologique des monophtongues est bien plus invasive en français que dans les deux autres langues.

L'approximation des diphtongues (cf. Table 3) peut être généralisée à travers l'échelle d'optimalité suivante : (1) diphtongue correspondante si elle est disponible ; (2) sinon, une séquence de deux articulations vocoïdes si elle est phonotactiquement acceptable ; (3) sinon, une monophtongue. Le français, n'ayant pas de diphtongues, fait appel aux catégories (2) et (3), l'allemand se sert des catégories (1) et (3), et le tchèque, des catégories (1) et (2). Ici encore, le français semble être plus invasif que les deux autres langues.

Anglais	Français	Allemand	Tchèque
eɪ ɔɪ aɪ əʊ aʊ eə	ɛ/e ɔj aj o au ɛʁ	eː ɔy ai oː au ɛːɐ	ɛj ɔj aj ɔu au ɛ(ː)r

TABLE 3 : Approximation des diphtongues dans l'échantillon. Les approximations des diphtongues [ʊə ɪə] ne sont pas attestées dans l'échantillon et [ɔɪ] est représenté par une seule entrée (*spoiler*). La restitution du /r/ dans la diphtongue [eə] est à la limite de la prononciation orthographique ; deux arguments peuvent cependant être soulevés en faveur de sa classification comme approximation phonologique : le /r/ est toujours prononcé en contexte de liaison, ainsi que dans les accents rhotiques de l'anglais.

Les règles d'approximation phonologique des consonnes sont répertoriées dans la Table 4. En considérant le nombre et le degré de transformations phonologiques, il semble que c'est le tchèque qui est légèrement plus invasif que les deux autres langues.

Anglais	Français	Allemand	Tchèque
p b t d k g ʧ ʤ m n f v s z ʃ r j l	sans modification	sans modification	sans modification
obstruantes en position finale (*gag, live*)	sans modification	dévoisement	dévoisement
/r/ potentiel en fin de syllabe (*mixer, airbag*)	ʁ	en principe, sans restitution	r
h	élision	sans modification	ɦ
ŋ	ŋ	ŋ	ŋ ; en fin de mot : ŋk
w	w	v	v

TABLE 4 : Approximation des consonnes dans l'échantillon. Le phonème /ʒ/ est absent de l'échantillon et les phonèmes /ð θ/ sont représentés chacun par une entrée. Nous faisons abstraction des différentes variantes du phonème /r/ dans les quatre langues.

Finalement, l'adaptation de l'accent se fait selon des règles assez transparentes : l'allemand, langue à accent libre, maintient l'accentuation d'origine, alors que le français et le tchèque, langues à accent fixe, régularisent sa position.

Les degrés d'invasivité dans les processus d'approximation phonologique peuvent être résumés de la manière suivante :

	Moins d'invasivité ⟷	Plus d'invasivité
monophtongues	All., Tch.	Fr.
diphtongues	All., Tch.	Fr.
consonnes	All., Fr.	Tch.
accentuation	All.	Tch., Fr.

L'allemand semble légèrement moins invasif que le tchèque, et bien moins invasif que le français. L'hypothèse 4 n'est donc pas infirmée, mais sa reformulation est nécessaire : c'est le français qui est plus invasif que les deux autres langues en ce qui concerne les processus d'approximation phonologique.

5 Conclusion

L'étude typologique que nous avons menée est une contribution originale à la phonologie des emprunts : nous avons formalisé les processus d'adaptation à travers un paradigme de principes avant d'appliquer ce paradigme à un échantillon représentatif et comparable d'anglicismes dans trois langues européennes. La question d'« invasivité phonologique » a été décomposée en quatre hypothèses, dont trois ont été confirmées.

Premièrement, l'approximation phonologique (substitution de phonèmes natifs aux phonèmes étrangers) se démarque comme le principe le plus vigoureux dans les trois langues analysées. Dans chacune d'entre elles, il se matérialise par des règles spécifiques de projection phonologique, déterminées par les différences qui existent entre le système phonologique de la langue donneuse et celui de la langue emprunteuse. Cette projection se caractérise par une grande régularité.

Deuxièmement, la prononciation orthographique (phonétisation des graphèmes) est assez fréquente en français, relativement fréquente en tchèque et peu fréquente en allemand. On pourrait spéculer sur les causes de ce phénomène : Le recours à ce type de prononciation est-il inversement proportionnel à la notoriété de l'anglais dans la population (dans l'histoire et aujourd'hui) ? S'agit-il de la « voie du moindre effort » dans une situation où l'approximation phonologique mène à un résultat qui est trop éloigné de la structure phonologique du mot dans la langue source?

Troisièmement, la prononciation authentique est une option assez fréquente pour l'adaptation des anglicismes en allemand, alors qu'elle est absente des dictionnaires français et tchèque. L'allemand semble donc particulièrement « perméable » aux transferts phonologiques directs de l'anglais, phénomène qui est conditionné par la proximité des systèmes phonologiques et catalysé par la notoriété générale de l'anglais.

Quatrièmement, les mécanismes d'approximation phonologique sont plus invasifs en français que dans les deux autres langues, et ce notamment en ce qui concerne le système vocalique. Malgré cela, les formes phonologiques adaptées restent identifiables et l'homophonie n'est pas un facteur qui compliquerait l'adoption des anglicismes.

Ces observations nous permettent de conclure que c'est l'allemand qui manifeste le moins d'invasivité phonologique vis-à-vis des anglicismes, alors que le français se situe à l'autre bout de l'échelle. Le tchèque occupe une position intermédiaire, tout en étant plus proche de l'allemand.

Remerciements

Le présent texte a été préparé dans le cadre du projet GAČR 16-06012S.

Références

CALABRESE A., WETZELS W.L., eds. (2009). *Loan Phonology*. John Benjamins.

DUBĚDA T., HAVLÍK M., JÍLKOVÁ L., ŠTĚPÁNOVÁ, V. (2014). Loanwords and Foreign Proper Names in Czech: A Phonologist's View. *Language Structure and Language Use. Proceedings of the Olomouc Linguistics Colloquium 2013*, 313–321.

Duden. Universalwörterbuch (1989). 2. Aufl., Mannheim – Wien – Zürich: Dudenverlag.

GÖRLACH M., ed. (2001). *A Dictionary of European Anglicisms. A Usage Dictionary of Anglicisms in Sixteen European Languages*. Oxford University Press.

HOGG R., DENISON, D (2008). *A History of the English Language*, Cambridge University Press.

JONES D., ROACH P., SETTER J., ESLING J. (2011). *Cambridge English Pronouncing Dictionary*. 18th edition. Electronic version. Cambridge University Press.

LANGE F. (2015). *Standardaussprache englischer Namen im Deutschen*. Berlin: Frank & Timme.

MUHVIĆ-DIMANOVSKI V. (1995). Anglicisms in German: the problem of variants. *Studia romanica et anglica Zagrabiensia 42*.

Nový akademický slovník cizích slov (2005). Praha: Academia. [Nouveau dictionnaire académique des emprunts]

Le Petit Robert de la langue française (2012). Version numérique. Dictionnaires Le Robert.

RETMAN R. (1978). L'adaptation phonétique des emprunts à l'anglais en français. *La Linguistique 14/I*, 111–124.

ROMPORTL M., ed. (1978). *Výslovnost spisovné češtiny*. Praha: Academia. [Prononciation du tchèque standard]

Investigation glottographique et laryngoscopique de la transition entre les deux principaux mécanismes laryngés

Arthur Givois[1], Didier Demolin[1], Lise Crevier-Buchman[1, 2], Angélique Amelot[1]
(1) Laboratoire de Phonétique et Phonologie, 19 rue des Bernardins 75005 Paris, France
(2) Hôpital Européen Georges Pompidou, 20 rue Leblanc 75015 Paris, France
arthur.givois@gmail.com, didier.demolin@univ-paris3.fr,
lise.buchman@numericable.fr, angelique.amelot@univ-paris3.fr,

RÉSUMÉ

Cet article étudie par une approche descriptive la transition entre le premier et le second mécanisme laryngé. Des mesures électroglottographiques ont été réalisées simultanément à des captures d'images par laryngoscopie sur deux sujets : une femme et un homme. Des différences de comportement entre les deux sujets ont été observées. Un mouvement vertical de grande amplitude du larynx est systématiquement observé au moment de la transition chez le sujet masculin, tandis que des modifications de petite amplitude de la distance entre paroi pharyngale et épiglotte, ou de la compression des plis aryépiglottiques sont remarquées chez le sujet féminin. Ces changements de configurations s'effectuent de façon continue chez cette dernière alors qu'un changement soudain de l'activité des plis vocaux a lieu à un instant précisément localisé pour les productions des deux sujets. Ces différences d'ajustements laryngés sont liées à des modifications des paramètres mécaniques dont dépendent la fréquence fondamentale et qui restent à estimer.

ABSTRACT

Glottographic and laryngoscopic investigation of the transition between the two main laryngeal mechanisms

This study focuses on the glottal source level and the transition between first and second laryngeal mechanism with a descriptive approach. Electroglottographic measurements and laryngoscopic images are recorded simultaneously on two subjects: a women and a men. Differences were noticeable in the behavior of the larynx: a high-amplitude movement of the vertical position is systematically visible at the transition instant for the male subject, whereas low-amplitude movements concerning the distance between pharyngeal wall and epiglottis, or compression change of the arytenoids because of a rotation are noticeable for the women. This subject realizes these modifications continuously, although sudden changes happen in the vocal folds activity at a precise instant related to the transition of mechanism for the two subjects. These differences of laryngeal adjustements are linked with modifications of mechanical parameters, whose fondamental frequency is dependant and which remain to estimate.

MOTS-CLÉS : Ajustements laryngés, glottographie, fréquence fondamentale, mécanismes laryngés

KEYWORDS: Laryngeal adjustements, glottography, fundamental frequency, laryngeal mechanisms

1 Introduction

Les différentes configurations du vibrateur laryngé pour produire des sons voisés ont conduit à distinguer quatre mécanismes de production au niveau de la source. Les critères de différenciation reposent sur la masse et la longueur vibrante des plis vocaux impliquées dans la vibration, l'action des muscles vocaux commandant les paramètres de tension et de masse vibrante, ainsi que des phénomènes de discontinuité liés aux transitions (Roubeau, 1993). Ces "mécanismes laryngés" – usuellement notés M0, M1, M2 et M3- sont listés par ordre croissant des fréquences fondamentales productibles dans chacun d'eux. Les deux mécanismes principalement utilisés par l'être humain en voix parlée sont le M1 et le M2. Les termes "chest" ou "modal voice" sont régulièrement utilisés dans la littérature pour désigner le M1 lorsqu'elles qualifient le comportement de la source glottique. Il en va de même avec le terme "falsetto" qui désigne le M2 (Roubeau et al., 2009). L'électroglottographie (EGG) est aujourd'hui la méthode non-invasive la plus fiable pour différencier les mécanismes laryngés : au niveau macroscopique, l'amplitude du signal EGG montre de plus grandes variations de surface d'accolement des plis vocaux en mécanisme M1 qu'en M2. Au niveau d'une période, les signaux EGG enregistrés simultanément à des images obtenues par cinématographie ultra-rapide ou photographie strobostropique ont permis d'associer l'évolution d'une onde de l'EGG aux différentes phases d'ouverture et de fermeture des plis vocaux. L'analyse de ces enregistrements coordonnés a montré que des évolutions plus ou moins rapides surviennent à l'échelle d'une période, la dérivée du signal EGG (dEGG) devient alors un outil pertinent pour caractériser les vibrations : les instants d'ouverture initiale et de fermeture complète de la glotte sont considérés proches des instants où des pics sont visibles sur le signal dEGG. Des procédures de détection automatique du quotient de contact ont été mises en place et ont permis d'étudier les tendances du comportement vibratoire des plis vocaux de façon quantitative avec des mesures non-invasives. Le quotient de contact est en moyenne 0.27 plus grand en M1 qu'en M2, il varie de 0.2 à 0.7 en M1 et de 0.05 à 0.5 en M2 (Henrich et al., 2005). La forme d'onde est plus symétrique en M2 car si la vitesse d'ouverture est inférieure à la vitesse de fermeture en M1, ces deux grandeurs sont proches de l'égalité en M2 (Roubeau et al., 2009).

Si le succès des mesures EGG s'explique en grande partie par son caractère non-invasif, sa principale limite repose sur l'impossibilité de décomposer l'évolution du contact des plis dans le plan horizontal (axe antéro-postérieur) et vertical (axe inférieur-supérieur). Ces informations pourraient potentiellement expliquer les différences entre les pics du signal dEGG et les instants d'ouverture initiale et de fermeture complète des plis, mises en évidence grâce à l'augmentation de la fréquence des images de la cinématographie (Orlikoff et al., 2012; Herbst et al., 2014). D'autre part, ces informations permettraient d'améliorer la compréhension des phénomènes de "double pics" du signal dEGG qui surviennent dans certains schémas de vibrations glottiques liés à des différences de phases d'ouverture ou de fermeture des plis vocaux le long de l'axe antéro-postérieur. Enfin, elles rendraient possible l'estimation de l'ouverture de la glotte et l'établissement d'un lien avec les modèles d'onde de débit glottique. La photoglottographie (PGG) est une technique qui permet également d'estimer l'ouverture de la glotte (Harden, 1975). Des analyses de signaux PGG sur des voyelles produites dans les deux principaux mécanismes laryngés mécanismes ont pu mettre en évidence des différences significatives au niveau de l'amplitude et de la forme de l'ouverture (Kitzing, 1982). Les opportunités offertes par cet outil n'ont pas été plus développées en raison de son caractère invasif. Un photoglottographe externe (ePGG) développé au Laboratoire de Phonétique et Phonologie de Paris 3 donne la possibilité d'acquérir des mesures PGG sans introduire de source de lumière ou de capteur photovoltaïque par voie orale ou nasale (Honda & Maeda, 2009). Ce nouvel outil n'a pour l'instant été utilisé que pour détecter des mouvements d'abduction et d'adduction totale de la glotte (Honda & Maeda, 2008). La robustesse du protocole ne permet pas à ce jour de mesurer les vibrations des plis vocaux avec cet appareil.

L'essentiel des études qui traitent le problème du contrôle neuro-musculaire de la fréquence fondamentale sont historiques et s'appuient sur des mesures électromyographiques (Hirano et al., 1970; Gay et al., 1972). Celles-ci ont montré que l'élévation de la F0 est corrélée avec l'activité du muscle cricothyroïdien (CT) sur toute la tessiture et indépendamment du mécanisme employé. La différence entre M1 et M2 repose principalement sur la contribution du muscle thyroaryténoïdien (TA), dont l'activité est liée à la masse vibrante (Hirano, 1982). Une limite physiologique de raideur peut être atteinte dans les hautes fréquences pour le M1 (Hirano et al., 1969; Titze, 1994). Une discontinuité de la F0 peut alors survenir en raison de l'arrêt soudain de l'activité du TA, alors que sa tension musculaire augmente graduellement (Švec et al., 1999). Un mouvement d'éloignement des bandes ventriculaires a lieu lors de la transition du M1 vers le M2, tandis qu'un rapprochement est relevé lors de la transition inverse (Bailly, 2009).

La corrélation de l'activité du CT avec la F0 est due aux mouvements de rotation et de translation du muscle cricothyroïdien qui provoquent une augmentation de la tension, dont l'effet est plus significatif que l'augmentation de la longueur. La montée du larynx contribue une augmentation supplémentaire de l'activité du muscle cricothyroïdien (Honda, 2004). Des images obtenues par laryngoscopie (Edmondson & Esling, 2006), imagerie par résonance magnétique (Echternach et al., 2010), ou des résultats tirés de signaux EGG (Henrich Bernardoni et al., 2014) montrent que la hauteur du larynx est en moyenne plus élevée chez les hommes en M2 qu'en M1. Il est rapporté dans ces études l'existence d'une grande variabilité inter-individuelle des positions du larynx en fonction des mécanismes. Ceci confirme que les stratégies neuromusculaires mises en œuvre pour le contrôle des paramètres acoustiques de la parole varient d'un mécanisme à l'autre. Le quotient ouvert dont la valeur dépend de ces stratégies est d'ailleurs corrélé différemment avec la hauteur et l'intensité suivant le mécanisme employé (Henrich et al., 2005). Le contrôle de la F0 en fonction des mécanismes résulte donc de la coordination complexe de l'activité musculaire laryngée qui est propre à chaque individu. L'expertise vocale et le sexe sont des facteurs qui augmentent cette variabilité. Les fréquences fondamentales de transition sont en moyenne inférieures chez les hommes par rapport aux femmes et les intervalles de ruptures sont plus élevés (Roubeau et al., 2009). Ces grandeurs dépendent des dimensions des organes du larynx, notamment de la longueur et de l'épaisseur des plis vocaux. Les rapports d'amplitude de l'EGG entre M1 et M2 sont également plus élevés pour les hommes que pour les femmes.

Cette étude propose de mettre en évidence cette variabilité des stratégies de contrôle de F0 entre individus en fonction des mécanismes. La description d'images laryngoscopiques associée à l'analyse de signaux EGG ont été réalisées sur un homme et une femme dans cette optique.

2 Protocole expérimental

Par son caractère invasif, le laryngoscope nécessite de réaliser les mesures dans un environnement hospitalier. Les expériences se sont déroulées dans un cabinet d'oto-rhino-laryngologie de l'Hôpital Européen Georges Pompidou.

2.1 Matériel

Un microphone électrostatique fixé sur un nasofibroscope placé à 30 cm des sujets a permis d'enregistrer les signaux acoustiques échantillonnés à 44100 Hz. Un électroglottographe a enregistré des signaux EGG échantillonnés à une même fréquence. Un nasofibroscope souple (Kay-Pentax FNL10RP3) équipé d'une lumière froide (halogène CLK-4, Olympus) dont l'extrémité est reliée à un système d'enregistrement vidéo (DigitalStrobe, RLS91000, Kay Elemetrics) ont permis d'enregistrer des images à une fréquence de 25 im/s et une résolution de 568 × 454 pixels. Ces enregistrements ont été effectués simultanément.

2.2 Locuteurs et tâches

Les sujets de l'expérience sont un homme et une femme musiciens mais non spécialistes en chant lyrique. Pour ces deux sujets, la transition est mise en évidence avec des glissandi ascendants et descendants sur différentes voyelles. Les changements de mécanismes du sujet masculin sont plus évidents à percevoir et à repérer. Ce sujet a en plus effectué des transitions sur une même fréquence fondamentale, une même hauteur et une même voyelle. Ces sujets ont produit différentes phonations - naturelle, criée, parlée, à différentes hauteurs fixées de leur tessiture - sur une phrase de référence (« Où a-t-il mis ses chaussettes »). La séquence alternant consonnes fricatives sourdes et voyelles afin de visualiser les mouvements d'abduction et d'adduction des plis vocaux sur les images et les signaux a été enregistrée. Une production chantée libre d'une durée approximative de dix secondes a également été demandée aux sujets.

2.3 Mesures

Les instants de fermetures et d'ouvertures glottiques ont été estimés par des méthodes de seuillage du signal EGG (deux seuils étant fixés à 35% et 50% de la différence entre le maximum et le minimum de la différence du signal EGG sur une période (Rothenberg & Mahshie, 1988), ainsi qu'à l'aide de la méthode Decom fondée sur la détection des pics de la dérivée du signal d'EGG (Henrich et al., 2004). Une quatrième méthode proposée dans l'étude d'Howard (1995) repose sur l'établissement d'un seuil fixé à 3/7 pour l'estimation de l'instant de l'ouverture et sur le pic du signal dEGG pour estimer l'instant de fermeture. Les fréquences fondamentales et les quotients de contacts estimés par ces quatre méthodes sont systématiquement comparés pour tous les enregistrements analysés. Les évolutions sont globalement similaires, les différences entre les valeurs restent le plus souvent constantes au cours d'une production pour les quatre méthodes de calcul. Des phénomènes locaux tels que des augmentations ou diminutions inexpliquées et brusques du quotient de contact surviennent dans certains cas. Les méthodes retenues dans les exemples présentés sont celles où n'apparaissent pas ces phénomènes inattendus. La méthode Decom est appliquée à la transition sur une même hauteur au glissando ascendant du sujet masculin, la méthode de Howard est appliquée sur les glissandi du sujet féminin, tandis que la méthode de seuillage à 35% est appliquée sur le glissando descendant du sujet masculin. Les discontinuités de la fréquence fondamentale et du quotient de contact estimées par ces méthodes sont avec les changements d'amplitudes de l'électroglottographe les principaux critères d'identification des transitions de mécanismes.
Les signaux EGG mesurés sur le sujet féminin ont été perturbés par un bruit qui n'a pas été visualisé lors de l'enregistrement. Une procédure de débruitage par analyse harmonique et synthèse additive a été entreprise sur la base du fait que l'information utile de l'EGG n'est contenue que dans la partie strictement harmonique du signal. Un filtre médian a également été appliqué sur les signaux EGG avant d'estimer les instants d'ouverture et de fermeture glottiques afin de pouvoir mieux détecter les pics d'ouverture sur les signaux dEGG. Ces estimations sont erronées dans les hautes fréquences des glissandi en raison d'une analyse et d'une reconstruction moins précises, aboutissant à des valeurs non interprétables des quotients de contact dans les hautes fréquences produites en M2.

Des mesures de longueurs en pixels ont également été effectuées sur les images laryngoscopiques. Pour chaque photographie analysée, trois distances notées L_1, L_2 et l_3 sont comparées conformément à l'exemple de la figure 1. L_1 et L_2 sont repérées par les distances entre la commissure inférieure des plis vocaux aux deux points supposés comme les plus hauts des plis aryépiglottiques, et l_3 désigne la distance entre les bords libres des bandes ventriculaires au niveau de la demi-longueur visible des plis vocaux. Les rapports L_1/l_3 et L_2/l_3 sont utilisés pour tenter de quantifier les changements d'inclinaison et la distance du larynx à la caméra.

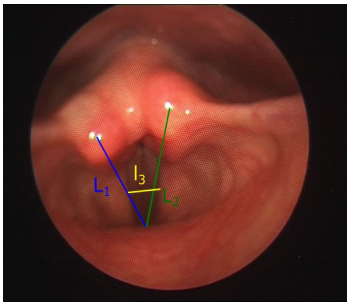


FIGURE 1: Représentation des longueurs mesurées sur un exemple d' image du larynx (sujet masculin)

N° d'image	L_1	L_2	l_3	L_1/l_3	L_2/l_3
1	221	251	67	3.30	3.74
2	237	285	79	3.01	3.63
3	203	238	72	2.82	3.29

Table I : Longueurs en pixels et rapports de longueurs sur les images enregistrées lors de la transition du M1 vers le M2 sur une hauteur tenue par le sujet masculin (figure 2).

3 Résultats

3.1 Sujet masculin

Les phénomènes de modification de l'amplitude du signal EGG et du quotient de contact se retrouvent sur l'ensemble des productions réalisées par le sujet masculin comme l'illustrent les trois exemples de transitions sur une même hauteur (figure 2) et sur des glissandi ascendant et descendant (figures 3 et 4). Des discontinuités de la F0 traduisent des phénomènes de perte de contrôle. En M2, l'amplitude de l'EGG est parfois trop faible pour pouvoir estimer la fréquence fondamentale et le quotient de contact.

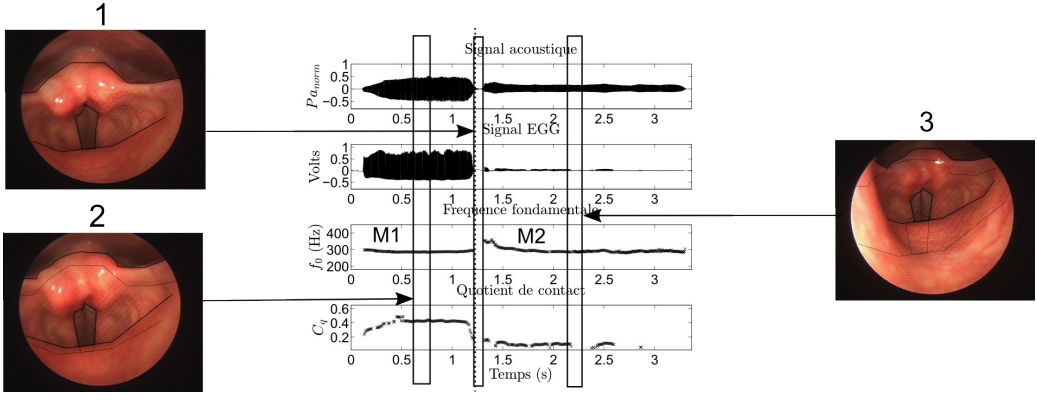


Figure 2 : Signaux acoustique et EGG mesurés, F0 et quotients de contact estimés lors d'une transition du mécanisme M1 vers le M2 sur une hauteur tenue par le sujet masculin. Voyelle [a], f0 = 289 Hz en moyenne. La barre verticale pointillée désigne le moment de la transition identifié sur les signaux. Les trois images capturées sont prises à un instant associé au M1, à la transition et un instant associé au M2. Le bord supérieur des plis aryépiglottiques (PA), le bord inférieur de l'union des PA et des bandes ventriculaires (BV) et les bords visibles des plis vocaux sont marqués en trait plein sur chaque image. Les traits pointillés désignent les mêmes frontières mais tirées de l'image précédente.

Lors d'une transition réalisée à F0 constante, l'évolution de la position et de la taille occupées par les différents organes sur les enregistrements vidéo-nasofibroscopiques traduit un mouvement descendant du larynx qui a lieu au moment de la transition (*cf* figure 2 et Table I). Le larynx monte légèrement vers l'avant entre les images 1 et 2, et descend brusquement entre les images 2 et 3. Une plus grande partie de l'épiglotte devient visible à l'image et masque alors la partie antérieure des PV

qui paraissent raccourcis en longueur. La largeur visible des PV varie peu par rapport à leur longueur, probablement en raison d'un éloignement des bandes ventriculaires (BV).

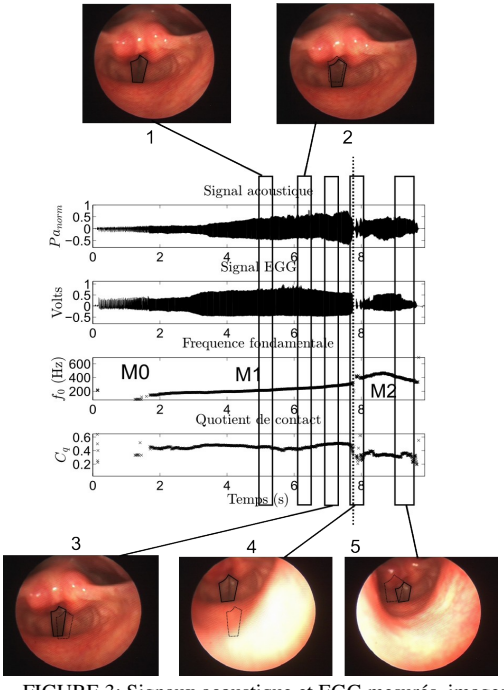


FIGURE 3: Signaux acoustique et EGG mesurés, images enregistrées, f0 et quotients de contact Cq estimés lors d'un glissando ascendant réalisé par le sujet masculin.

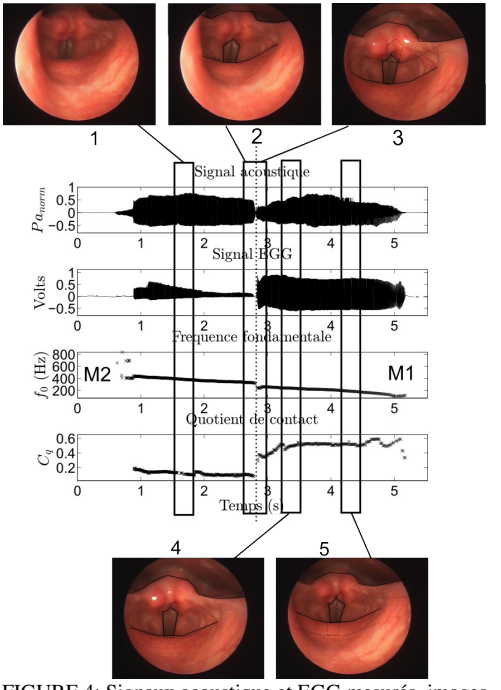


FIGURE 4: Signaux acoustique et EGG mesurés, images enregistrées, f0 et quotients de contact estimés lors d'un glissando descendant réalisé par le sujet masculin.

L'analyse des images enregistrées lors d'un glissando ascendant (*cf* figure 3 et Table II) produit par le même sujet permet de constater ces mêmes phénomènes de descente et de basculement du larynx lors du passage du M1 au M2. Trois images ont été extraites de la partie produite en M1 pour illustrer le mouvement lent des organes du larynx en comparaison avec l'ample changement de position qui se produit au moment de la transition. Ces trois images montrent une ascension du larynx avec la F0. La sensation d'effort vocal se traduit par un serrage supraglottique auquel semblent participer les BV. Une constriction du vestibule laryngé survient au moment de la transition alors que le larynx est toujours en position haute (image 4) et précède le mouvement descendant et le basculement du larynx vers l'arrière qui caractérise la transition (image 5).

N° d'image	L_1	L_2	l_3	L_1/l_3	L_2/l_3
1	239	262	67	3.58	3.93
2	226	258	67	3.39	3.87
3	257	274	66	3.89	4.15
4	212	224	79	2.69	2.84
5	181	201	63	2.88	3.20

Table II : Longueurs en pixels et rapports de longueurs sur les images enregistrées lors du glissando ascendant (figure 3) réalisé par le sujet masculin.

N° d'image	L_1	L_2	l_3	L_1/l_3	L_2/l_3
1	198	235	67	2.96	3.51
2	196	237	68	2.88	3.47
3	217	243	67	3.24	3.63
4	205	234	62	3.31	3.79
5	188	214	68	2.81	3.20

Table III : Longueurs en pixels et rapports de longueurs sur les images enregistrées lors du glissando descendant (figure 4) réalisé par le sujet masculin.

La stratégie concernant la position du larynx semble être employée indépendamment du sens de la mélodie. Le glissando descendant (figure 4 et Table III) illustre le même mécanisme réalisé en sens inverse. Le larynx est en position basse en M2 (image 1), les PV occupent une faible proportion sur les images par rapport à l'épiglotte en raison de la constriction du vestibule laryngé et de leur éloignement par rapport à la caméra. Le larynx effectue un mouvement ascendant en M2 comme l'illustre la légère modification de la position postérieure des PV de l'image 2. Il bascule ensuite vers l'avant en maintenant le serrage épilaryngé au moment de la transition (image 3). L'image 4 fait apparaître un changement des positions au niveau des extrémités des PA par rapport à l'image 3. Un mouvement descendant du larynx semble avoir lieu pendant la production des basses fréquences du M1 (image 5). Roubeau (1993) et Gay et al. (1972) suggèrent que ceci est dû à l'action des muscles extrinsèques du larynx.

3.2 Sujet féminin

Les glissandi produits par le sujet féminin montrent des discontinuités dans la fréquence fondamentale et pour le quotient de contact, en particulier dans le cas du glissando ascendant : il est donc supposé qu'une transition de mécanisme a lieu à ces instants. Les images et mesures présentées sur les figures 4 et 5 décrivent des variations plus continues de la disposition de l'appareil laryngé par rapport au sujet masculin.

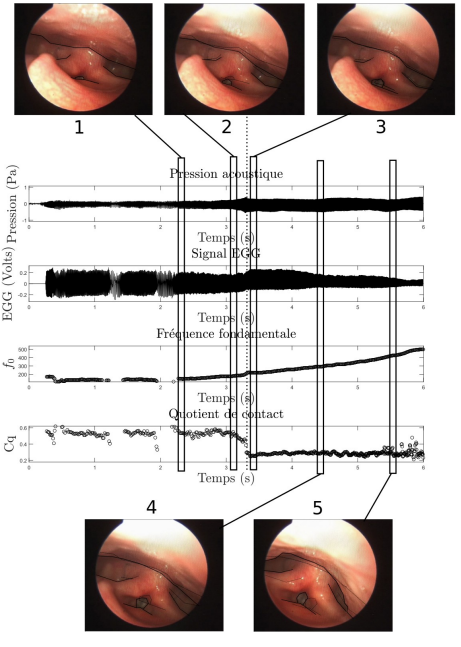


FIGURE 5: Signaux acoustique et EGG mesurés, images enregistrées, f0 et Cq estimés lors d'un glissando ascendant sur la voyelle [u] par le sujet féminin. Les discontinuités des signaux lors des temps t < 2.3s sont dus à transitions entre mécanismes M1 et M0.

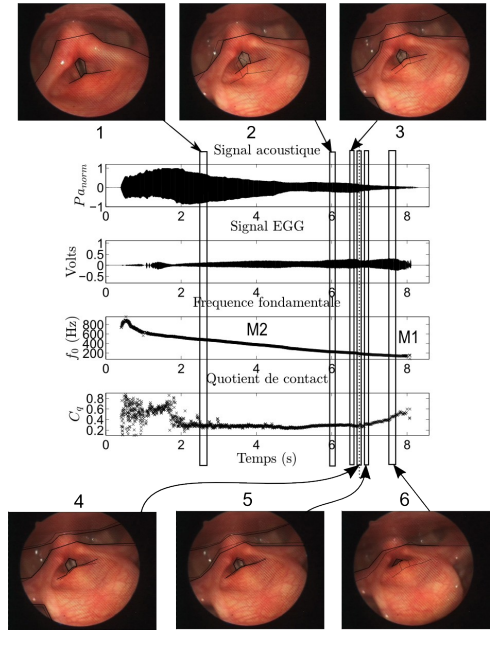


FIGURE 6: Signaux acoustique et EGG mesurés, images enregistrées, f0 et Cq estimés lors d'un glissando descendant sur la voyelle [u] effectué par le sujet féminin. Ces estimations sont erronées pour t < 2s environ.

Les images 1 à 6 de la figure 5 extraites du glissando descendant montrent que la surface visible des PV diminue de façon progressive en raison du recouvrement de l'épiglotte qui prend une place croissante dans la succession des images, comme le montrent les distances de la Table IV. Ces variations de parties visibles occupées par les PV et l'épiglotte sont dues à une descente progressive

du larynx. L'enchaînement entre les images 3 et 4, où la transition a lieu, se distingue par un rapprochement entre la paroi pharyngale et l'épiglotte. Les deux organes effectuent à ce moment un mouvement l'un vers l'autre, donnant la sensation d'un serrage laryngé tandis que les positions des PA et PV varient peu au cours de la succession entre les images 3, 4 et 5.

L'angle de vue proposé par les images tirées d'un glissando ascendant (figure 6 et Table V) confirme qu'une montée et un basculement du larynx s'effectuent progressivement au cours de la production, particulièrement dans les basses fréquences (*cf* images 1 et 2). Une augmentation de la surface visible des PV et un mouvement d'éloignement entre la paroi pharyngale et l'épiglotte s'effectuent également de façon continue, l'épiglotte n'est en effet presque plus visible à partir de l'image 4. Cependant, ces deux mouvements sont réalisés différemment au moment de la transition par rapport aux autres instants. La position visible des limites de la paroi pharyngale ne varie quasiment pas entre les images 2 et 3 par rapport aux autres enchaînements (les limites sur l'image 5 peuvent être mal interprétées en raison d'un changement de distance focale). L'augmentation de la surface visible des PV s'effectue entre les images 2 et 3 par l'éloignement des PA au niveau de la commissure postérieure des PV.

N° d'image	L_1	L_2	l_3	L_1/l_3	L_2/l_3
1	137	157	45	3.0	3.45
2	144	153	43	3.32	3.51
3	119	128	54	2.18	2.35
4	121	136	52	2.31	2.60
5	90	95	48	1.90	1.98
6	72	75	56	1.30	1.34

Table IV : Longueurs en pixels et rapports de longueurs sur les images enregistrées lors du glissando descendant (figure 5) réalisé par le sujet féminin.

N° d'image	L_1	L_2	l_3	L_1/l_3	L_2/l_3
1	71	101	39	1.82	2.59
2	84	115	45	1.87	2.55
3	100	122	53	1.89	2.30
4	130	145	60	2.17	2.42
5	145	155	62	2.34	2.5

Table V : Longueurs en pixels et rapports de longueurs sur les images enregistrées lors du glissando ascendant (figure 6) réalisé par le sujet féminin.

4 Conclusion et perspectives

Dans cette étude, la comparaison de signaux EGG avec des images du larynx a montré que les deux sujets de l'étude peuvent utiliser différents moyens de contrôle de la fréquence fondamentale pour effectuer des tâches similaires, comme l'ont illustré les exemples de glissandi. Les mouvements verticaux du larynx constituent un marqueur essentiel du changements de mécanisme pour le sujet masculin, tandis que d'autres événements de plus faible amplitude - tels que les mouvements de rapprochements et d'éloignement entre l'épiglotte et la paroi pharyngale, ou une rotation des aryténoïdes - semblent être les principaux événements se produisant lors de la transition chez le sujet féminin. Cette étude illustre par quelques exemples le problème de la variabilité des stratégies musculaires et dispositions physiologiques employées par rapport à la répétabilité des observations réalisées au niveau de la source - modification de la forme d'onde et de l'amplitude de l'EGG, de masse et de longueur vibrante. Les images enregistrées ne permettent pas d'accéder à des informations sur la longueur vibrante, la partie inférieure des plis étant recouverte par l'épiglotte. Ainsi les variations des paramètres mécaniques de contrôle qui s'effectue manifestement de façon différente d'un sujet à l'autre n'ont pu être identifiées. La notion de contrôle des mécanismes et des comparaisons de transitions contrôlées et non-contrôlées n'a pas été évoquée et constitue une perspective possible à cette étude.

Références

BAILLY L. (2009). Interaction entre cordes vocales et bandes ventriculaires en phonation : exploration in-vivo, modélisation physique, validation in-vitro. PhD thesis, Université du Maine.

ECHTERNACH M., SUNDBERG J., MARKL M. & RICHTER B. (2010). Professional opera tenor's vocal tract configurations in registers. Folia Phoniatrica et Logopaedica, 62(6), 278–287.

EDMONDSON J. & ESLING J. (2006). The valves of the throat and their functioning in tone, vocal register and stress : laryngoscopic case studies. Phonology, 23(02), 157–191.

GAY T., HIROSE H., STROME M. & SAWASHIMA M. (1972). Electromyography of the intrinsic laryngeal muscles during phonation. Annals of Otology, Rhinology and Laryngology, 81(3), 401–409.

HARDEN R. (1975). Comparison of glottal area changes as measured from ultrahigh-speed photographs and photoelectric glottographs. Journal of Speech, Language, and Hearing Research, 18(4), 728–738.

HENRICH N., D'ALESSANDRO C., DOVAL B. & CASTELLENGO M. (2004). On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. The Journal of the Acoustical Society of America, (3), 1321–1332.

HENRICH N., D'ALESSANDRO C., DOVAL B. & CASTELLENGO M. (2005). Glottal open quotient in singing : Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency. The Journal of the Acoustical Society of America, (3), 1417–1430.

HENRICH BERNARDONI N., SMITH J. & WOLFE J. (2014). Vocal tract resonances in singing : Variation with laryngeal mechanism for male operatic singers in chest and falsetto registers. The Journal of the Acoustical Society of America, 135(1), 491–501.

HERBST C., LOHSCHELLER J., ŠVEC J., HENRICH N., WEISSENGRUBER G. & FITCH W. (2014). Glottal opening and closing events investigated by electroglottography and super-high-speed video recordings. The Journal of experimental biology, 217(6), 955–963.

HIRANO M. (1982). The role of the layer structure of the vocal fold in register control. Vox Humana. University of Jyvaskyla, p. 50Á.

HIRANO M., OHALA J. & VENNARD W. (1969). The function of laryngeal muscles in regulating fundamental frequency and intensity of phonation. Journal of Speech, Language, and Hearing Research, 12(3), 616–628.

HIRANO M., VENNARD W. & OHALA J. (1970). Regulation of register, pitch and intensity of voice. Folia Phoniatrica et Logopaedica, 22(1), 1–20.

HONDA K. (2004). Physiological factors causing tonal characteristics of speech : from global to local prosody. In Speech Prosody 2004, International Conference.

HONDA K. & MAEDA S. (2008). Glottal-opening and airflow pattern during production of voiceless fricatives : a new non-invasive instrumentation. The Journal of the Acoustical Society of America, 123(5), 3738–3738.

HONDA K. & MAEDA S. (2009). Procédé et équipement non-invasif de photoélectroglottographie. WO Patent App. PCT/FR2008/000,838.

HOWARD D. M. (1995). Variation of electrolaryngographically derived closed quotient for trained and untrained adult female singers. Journal of Voice, 9(2), 163–172.

KITZING P. (1982). Photo-and electroglottographical recording of the laryngeal vibratory pattern during different registers. Folia phoniatrica, 34, 234–241.

ORLIKOFF R., GOLLA M. & DELIYSKI D. (2012). Analysis of longitudinal phase differences in vocal-fold vibration using synchronous high-speed videoendoscopy and electroglottography. Journal of Voice, 26(6), 816–e13.

ROTHENBERG M. & MAHSHIE J. J. (1988). Monitoring vocal fold abduction through vocal fold contact area. Journal of Speech, Language, and Hearing Research, 31(3), 338–351.

ROUBEAU B. (1993). Mécanismes vibratoires laryngés et contrôle neuro-musculaire de la fréquence fondamentale. PhD thesis, Université Paris-Orsay, France.

ROUBEAU B., HENRICH N. & CASTELLENGO M. (2009). Laryngeal vibratory mechanisms : The notion of vocal register revisited. Journal of Voice, 23(4), 425–438.

ŠVEC J., SCHUTTE H. & MILLER D. (1999). On pitch jumps between chest and falsetto registers in voice : Data from living and excised human larynges. The Journal of the Acoustical Society of America, 106(3), 1523–1531.

TITZE I. (1994). Principles of voice production. National Center for Voice and Speech.

Modélisation bayésienne de la planification motrice des gestes de parole : Évaluation du rôle des différentes modalités sensorielles

Jean-François Patri[1,2,3,4] Julien Diard[3,4] Pascal Perrier[1,2]
(1) Univ. Grenoble Alpes, Gipsa-lab, F-38000 Grenoble, France
(2) CNRS, Gipsa-lab, F-38000 Grenoble, France
(3) Univ. Grenoble Alpes, LPNC, F-38000 Grenoble, France
(4) CNRS, LPNC, F-38000 Grenoble, France
Jean-Francois.Patri@gipsa-lab.grenoble-inp.fr,
Julien.Diard@upmf-grenoble.fr, Pascal.Perrier@gipsa-lab.grenoble-inp.fr

RÉSUMÉ

La prise en compte des informations auditives et proprioceptives dans le contrôle de la parole est mise en évidence par un nombre croissant de résultats expérimentaux. Cependant, les modèles de production imposent le plus souvent l'une ou l'autre des modalités, ou n'offrent pas de cadre formel pour évaluer leurs contributions respectives. Nous proposons d'explorer le rôle de ces modalités sensorielles dans la planification des gestes de parole à partir d'un modèle bayésien représentant la structure des connaissances mises en jeu dans cette tâche. Le modèle permet d'envisager trois mécanismes de planification, reposant sur la modalité auditive, proprioceptive ou sur les deux conjointement. Nous comparons des simulations obtenues par les deux premiers mécanismes de planification. Les résultats indiquent des réalisations articulatoires différentes mais donnant néanmoins des réalisations auditives qualitativement similaires dans leur variabilité.

ABSTRACT

Bayesian modeling of speech gesture motor planning: Evaluating the role of different sensory modalities

An increasing number of experimental results have identified a clear role of auditory and somatosensory information in speech motor control. However, most of the speech production models consider only one of these sensory modalities, or do not provide the possibility to formally evaluate the respective contribution of these modalities. We propose to explore the role of auditory and proprioceptive representations in speech gesture planning, based on a Bayesian model representing the structure of knowledge involved. The model allows to consider three planning mechanisms, based on the auditory or proprioceptive modality or the combination of both. We compare simulations obtained from the two first planning mechanisms. Results indicate differences in the generated articulatory patterns, giving rise however to qualitatively similar patterns of auditory variability.

MOTS-CLÉS : Contrôle moteur de la parole – Modélisation bayésienne – Multimodalité .

KEYWORDS: Speech motor control – Bayesian modeling – Multimodality.

1 Introduction

La remarquable capacité d'adaptation des gestes de parole à différents contextes et perturbations démontre que leur contrôle n'est pas le résultat d'un simple apprentissage stéréotypé mais met en jeu des mécanismes de planification, qui se révèlent en particulier par des stratégies d'anticipation (Noiray *et al.*, 2011). Quelles connaissances sont mises en jeu dans ces mécanismes et comment sont-elles structurées ? De nombreuses études expérimentales ont mis en évidence la prise en compte des information auditives et somato-sensorielles dans le contrôle et l'adaptation des gestes de parole (Lametti *et al.*, 2012). Le poids spécifique de ces retours sensoriels dans la la caractérisation des buts physiques associés aux unités phonologiques n'est pas connu et fait l'objet de nombreux débats, allant jusqu'à des enjeux importants en phonologie (Browman & Goldstein, 1992). Comment sont-ils pris en compte ? Ont-ils le même statut vis-à-vis du contrôle ? Sont-ils pris en compte de façon indépendante ou en combinaison ?

Cependant, à l'exception de DIVA (Guenther *et al.*, 2006), la plupart des modèles de production définissent les buts de parole dans l'une ou l'autre seulement de ces modalités sensorielles (Saltzman & Munhall, 1989; Stevens, 1993) et n'offrent pas de cadre formel pour une évaluation de leurs contributions respectives. Ce travail propose une méthodologie de modélisation pour étudier ces questions dans la planification des gestes de parole à partir d'un modèle bayésien des connaissances mises en jeu dans cette tâche (Bessière *et al.*, 2013). Ce modèle, qui est une extension d'un modèle précédent (Patri *et al.*, 2015), permet de définir et évaluer trois mécanismes de planification, reposant soit sur la modalité auditive, sur la modalité proprioceptive ou sur les deux conjointement.

Nous considérons la question de la planification des commandes motrices envoyées aux muscles de la langue pour la production de phonèmes isolés. Les six muscles les plus importants pour expliquer les déformations de la langue dans le plan sagittal sont représentés dans un modèle biomécanique bi-dimensionnel de la langue (Perrier *et al.*, 2003). Le niveau d'activation de chaque muscle est contrôlé selon la théorie du point d'équilibre par une variable de contrôle λ qui spécifie une longueur seuil pour le muscle, au delà de laquelle une force musculaire active est générée (Feldman, 1986). Les caractéristiques spectrales du signal acoustique correspondant à chaque configuration de langue sont obtenues par le calcul de la fonction d'aire, puis par l'utilisation d'un modèle harmonique de la propagation des ondes dans le conduit vocal (Badin & Fant, 1984). Nous considérons les trois premiers formants comme la représentation auditive des signaux acoustiques à produire (Paliwal *et al.*, 1983). La nature précise des indices sensoriels caractérisant la proprioception de la langue est encore mal connue. Nous supposons qu'ils sont caractérisés par les longueurs des fibres des six muscles considérés. Nous supposons que le cerveau dispose de deux modèles internes lui permettant de prédire les images auditives et proprioceptives correspondant à chaque configuration de paramètres de contrôle λ (Callan *et al.*, 2004). Nous supposons également que le cerveau dispose d'une représentation des buts auditifs et proprioceptifs associés à chaque phonème et de leurs domaines de variabilité compatibles avec une production correcte du phonème (Hickok, 2014). Ceci caractérise les distributions des images auditives et proprioceptives attendues pour chaque phonème. Nous supposons finalement que le cerveau est en mesure de comparer les prédictions sensorielles effectuées par les modèles internes moteurs avec les attentes correspondant aux caractérisations sensorielles des phonèmes (Blakemore *et al.*, 2000).

La Section 2 présente la construction du modèle, par une traduction de ces hypothèses dans le formalisme bayésien. La Section 3 expose les résultats de simulations de production de phonèmes isolés selon la modalité sensorielle considérée.

2 Présentation du modèle

2.1 Description

Variables probabilistes Chaque grandeur intervenant dans la planification des commandes motrices est associée à une variable probabiliste. Ces variables sont les suivantes :

$\boldsymbol{M}$ est la variable motrice, définie comme une variable 6-dimensionnelle : $M \equiv (\lambda_1, \ldots, \lambda_6)$. Les λ_i sont les variables de contrôle pilotant la contraction des six muscles de la langue considérés par le modèle biomécanique.

$\boldsymbol{\Phi}$ est une variable discrète composée des phonèmes { /i/, /e/, /ɛ/, /a/, /oe/, /ɔ/, /k/ } [1].

$\boldsymbol{S_a^M}$ et $\boldsymbol{S_a^\Phi}$ représentent les informations auditives mises en jeu dans la planification. S_a^M est générée à partir de la variable motrice M via le modèle interne auditif. S_a^Φ correspond à la représentation des buts auditifs associés à chaque phonème. Ce sont des variables tridimensionnelles continues caractérisées par les valeurs des trois premiers formants du signal acoustique : $S_a^{M,\Phi} \equiv (F_1, F_2, F_3)$.

$\boldsymbol{S_p^M}$ et $\boldsymbol{S_p^\Phi}$ représentent les informations proprioceptives mises en jeu dans la planification. S_p^M et S_p^Φ sont respectivement associées à M et Φ via le modèle interne et la représentation des buts proprioceptifs de chaque phonème. On définit alors S_p^M et S_p^Φ comme deux variables continues 6-dimensionnelles $S_p^{M,\Phi} \equiv (L_1, \ldots, L_6)$, où L_i est la longueur spécifiée du muscle i.

Les domaines des variables M, $S_{a,p}^M$ et $S_{a,p}^\Phi$ sont définis par les domaines de variation admissibles dans le modèle biomécanique.

$\boldsymbol{C_a}$ et $\boldsymbol{C_p}$ sont deux variables introduites pour contraindre la cohérence entre les valeurs prises respectivement par les variables S_a^M et S_a^Φ d'une part et S_p^M et S_p^Φ d'autre part. Ce sont des variables binaires égales à 1 ou 0 selon que les valeurs prises par S_a^M et S_a^Φ et S_p^M et S_p^Φ sont identiques ou pas. Lorsqu'elles sont fixées à 1 elles imposent la cohérence des prédictions motrices $S_{a,p}^M$ et des attentes $S_{a,p}^\Phi$ liées à la caractérisation des buts associés aux phonèmes.

Décomposition Le schéma de la Figure 1 décrit l'organisation des connaissances proposée dans notre modèle. Il permet de formuler la distribution de probabilité conjointe de l'ensemble des variables $P(M\ S_a^M\ S_p^M\ S_a^\Phi\ S_p^\Phi\ C_a\ C_p)$ par la décomposition suivante :

$$
\begin{aligned}
&P(M\ S_a^M\ S_p^M\ S_a^\Phi\ S_p^\Phi\ C_a\ C_p) \qquad (1)\\
&= \quad P(M)P(S_p^M \mid M)P(S_a^M \mid M)P(\Phi)P(S_p^\Phi \mid \Phi)P(S_a^\Phi \mid \Phi)P(C_a \mid S_a^M\ S_a^\Phi)P(C_p \mid S_p^M\ S_p^\Phi).
\end{aligned}
$$

Cette décomposition traduit, par ses hypothèses simplificatrices, les relations supposées entre variables. Par exemple, le terme $P(S_a \mid M)$ traduit une hypothèse d'indépendance entre S_a et S_p conditionnellement à la connaissance de M, qui est suggérée par l'hypothèse du modèle interne moteur, d'après lequel M seul suffit à caractériser S_a. En d'autres termes, S_p n'apporte aucune information supplémentaire à l'identification de S_a si M est connue.

[1] Cette liste est limitée aux phonèmes réalisables par le modèle biomécanique de la langue, sans usage spécifique des lèvres, non représentées dans cette version du modèle.

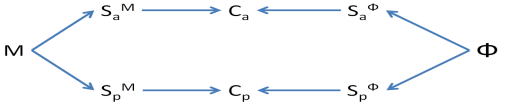


FIGURE 1 : Réseau représentant la décomposition de l'Équation (1).

Formes paramétriques Les termes de la décomposition de l'Équation (1) sont spécifiés par les formes paramétriques données ci-dessous.

$\boldsymbol{P(M)}$ représente la connaissance a priori sur la variable motrice. Nous supposons que toutes les configurations motrices utilisées pour la parole sont aisément réalisables et que de ce fait aucune n'est a priori privilégiée. $P(M)$ est donc identifiée à une distribution uniforme sur l'ensemble de son domaine de définition.

$\boldsymbol{P(S_a^M \mid M)}$ et $\boldsymbol{P(S_p^M \mid M)}$ représentent les connaissances des informations auditives et proprioceptives associées à la connaissance de M. Ces connaissances sont attribuées aux modèles internes prédisant les informations sensorielles à partir des commandes motrices. Notons $s_a^*(m)$ et $s_p^*(m)$ les valeurs prédites à partir de la valeur m de M. Les formes des distributions de probabilité associées sont définies par $P([S_{a,p}^M = s_{a,p}] \mid [M = m]) = \delta_{s^*_{a,p}(m)}(s_{a,p})$. δ_x correspond à la distribution Dirac centrée en x, traduisant le déterminisme, chez un locuteur donné, des relations entre les commandes motrices et le son produit. $P([S = s_{a,p}] \mid [M = m])$ est donc nulle tant que $s_{a,p} \neq s^*_{a,p}(m)$.

$\boldsymbol{P(S_a^\Phi \mid \Phi)}$ et $\boldsymbol{P(S_p^\Phi \mid \Phi)}$ représentent les connaissances des buts sensoriels correspondant à un phonème donné. On les identifie aux distributions des signaux sensoriels observés expérimentalement pour chaque phonème. Les productions acoustiques de chaque phonème étant caractérisées par des ellipsoïdes de dispersion dans l'espace des trois premiers formants, nous identifions $P(S_a^\Phi \mid \Phi)$ à une loi normale spécifiée par les paramètres de l'ellipsoïde correspondant au phonème considéré. La caractérisation des signaux proprioceptifs correspondant aux phonèmes est plus délicate. Nous supposons que la caractérisation auditive des phonèmes se met d'abord en place, permettant de guider les premières productions du sujet. La prise en compte des informations proprioceptives associées à ces productions initiales $P(S_p^M \mid \Phi\ [C_a = 1])$ permettrait alors au sujet de construire peu à peu une caractérisation proprioceptive de ses phonèmes $P(S_p^\Phi \mid \Phi)$. En d'autres termes, cet apprentissage identifie $P(S_p^\Phi \mid \Phi)$ au résultat de l'inférence $P(S_p^M \mid \Phi\ [C_a = 1])$:

$$P(S_p^\Phi \mid \Phi) \equiv P(S_p^M \mid \Phi\ [C_a = 1]) \propto \sum_m P(S_p^M \mid [M = m]) P([S_a^\Phi = s_a^*(m)] \mid \Phi). \tag{2}$$

$\boldsymbol{P(C_a \mid S_a^M\ S_a^\Phi)}$ et $\boldsymbol{P(C_p \mid S_p^M\ S_p^\Phi)}$ correspondent enfin aux contraintes de cohérence :

$$P([C_{a,p} = 1] \mid [S_{a,p}^M = s^M]\ [S_{a,p}^\Phi = s^\Phi]) = \begin{cases} 1 & \text{si } s^M = s^\Phi \\ 0 & \text{sinon.} \end{cases} \tag{3}$$

2.2 Utilisation du modèle

Nous simulons le mécanisme par lequel le cerveau sélectionnerait, selon nos hypothèses, les commandes motrices à envoyer aux muscles de la langue pour produire un phonème. Nous étudions donc l'inférence de la variable M à partir du phonème que l'on cherche à produire. Le modèle nous permet d'aborder cette question de trois façons : a) en imposant avec $C_a = 1$ la cohérence des variables

auditives S_a^M et S_a^Φ, l'inférence de $P(M \mid \Phi\ [C_a = 1])$ planifie les commandes motrices pour la production d'un phonème reposant sur les connaissances auditives seules ; b) de même, en imposant $C_p = 1$, $P(M \mid \Phi\ [C_p = 1])$ planifie les commandes motrices reposant sur les connaissances proprioceptives seules ; c) finalement en imposant à la fois $C_a = 1$ et $C_p = 1$, $P(M \mid \Phi\ [C_p = 1][C_a = 1])$ planifie les commandes motrices à partir des connaissances auditives et proprioceptives, suggérant une caractérisation bimodale des buts moteurs de parole.

Ces trois inférences sont obtenues à partir de la distribution de probabilité conjointe $P(M\ S_a^M\ S_p^M\ S_a^\Phi\ S_p^\Phi\ C_a\ C_p)$ spécifiée par l'Équation (1). Pour $P(M \mid \Phi\ [C_a = 1])$ on a :

$$P(M|\Phi[C_a = 1]) = \frac{P(M\Phi[C_a = 1])}{P(\Phi[C_a = 1])} = \frac{\sum_{S_a^M, S_a^\Phi, S_p^M, S_p^\Phi, C_p} P(M S_a^M S_p^M S_a^\Phi S_p^\Phi [C_a = 1] C_p)}{\sum_{M, S_a^M, S_a^\Phi, S_p^M, S_p^\Phi, C_p} P(M S_a^M S_p^M S_a^\Phi S_p^\Phi [C_a = 1] C_p)}.$$

En remplaçant $P(M\ S_a^M\ S_p^M\ S_a^\Phi\ S_p^\Phi\ C_a\ C_p)$ par l'expression de l'Équation (1) et en effectuant les sommations pour le cas d'une valeur m de M on a :

$$P([M = m] \mid \Phi\ [C_a = 1]) \quad \propto \quad P([S_a^\Phi = s_a^*(m)] \mid \Phi), \tag{4}$$

où le symbole de proportionnalité $\propto$ tient compte du facteur de normalisation. De façon similaire on obtient :

$$P([M = m] \mid \Phi\ [C_p = 1]) \quad \propto \quad P([S_p^\Phi = s_p^*(m)] \mid \Phi), \tag{5}$$

$$P([M = m] \mid \Phi\ [C_a = 1]\ [C_p = 1]) \quad \propto \quad P([S_a^\Phi = s_a^*(m)] \mid \Phi) P([S_p^\Phi = s_p^*(m)] \mid \Phi). \tag{6}$$

3 Résultats

3.1 Comparaison formelle des inférences auditive et proprioceptive

Avant d'exposer les résultats de l'implémentation du modèle, notons un premier résultat formel. Les Équations (2) et (5) donnent lieu à :

$$P([M = m] \mid \Phi\ [C_p = 1]) \quad \propto \quad \sum_{m'} P([S_p^M = s_p^*(m) \mid [M = m']) P([S_a^\Phi = s_a^*(m')] \mid \Phi). \tag{7}$$

En comparant à l'Équation (4), on constate que $P([M = m] \mid \Phi\ [C_p = 1]) \neq P([M = m] \mid \Phi\ [C_a = 1])$. Autrement dit, les branches auditive et proprioceptive ne proposent pas les mêmes inférences motrices. Cela peut paraitre étonnant étant donné que la caractérisation proprioceptive des phonèmes a été définie sur la base de leur caractérisation auditive. On peut remarquer que les inférences seraient identiques si chaque valeur de S_p^M était générée par une unique valeur de M. En effet dans ce cas la somme sur m' dans l'Équation (7) se réduirait à un unique terme : celui correspondant à la valeur m considérée pour M. Dans ce cas là on obtiendrait :

$$P([M = m] \mid \Phi\ [C_p = 1]) \propto P([S_a^\Phi = s_a^*(m)] \mid \Phi) \propto P([M = m] \mid \Phi\ [C_p = 1]). \tag{8}$$

Les branches proprioceptive et auditive deviendraient alors équivalentes. Cependant, étant donné qu'une même configuration articulatoire de la langue peut être obtenue à partir de différentes coactivations musculaires, l'équivalence précédente n'est pas assurée à priori. Ce serait également le cas

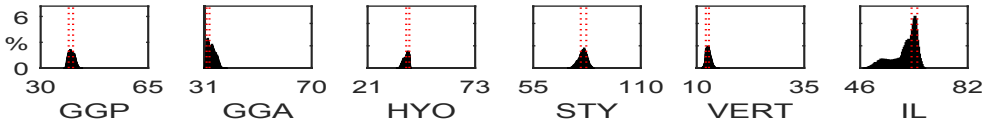


FIGURE 2 : Histogrammes des échantillons de S_p^Φ correspondant à l'inférence de l'Équation (2). Les lignes verticales indiquent la position des noyaux gaussiens retenus pour l'approximation de ces distributions.

si chaque valeur de S_p^M avait exactement le même nombre d'antécédents dans l'espace de M, mais aucune observation expérimentale ne laisse penser que cela puisse être le cas, et ceci est confirmé par notre modèle biomécanique.

Jusqu'à présent nous avons supposé une identification exacte de $P(S_p^\Phi \mid \Phi)$ avec l'inférence de $P(S_p^M \mid \Phi\,[C_a = 1])$ donnée par l'Équation (2). Cette identification formelle caractérise la distribution des valeurs de S_p^Φ dont les valeurs de S_a associées satisfont les cibles auditives. Bien que nous ayons choisi des cibles auditives simples, caractérisées par des gaussiennes, rien ne garantit que la distribution correspondante des valeurs de S_p^Φ soit également simple. La Figure 2 illustre les histogrammes des longueurs des muscles correspondant aux distribution marginales des échantillons de S_p obtenus par l'identification de l'Équation (2) obtenue par échantillonnage de type Monte-Carlo par chaines de Markov (MCMC) pour le phonème /e/. Bien que ces distributions marginales se caractérisent dans leur ensemble par des pics localisés, on observe également une structure plus complexe (en particulier pour le muscle Inferieur Longitudinalis dans l'exemple de la Figure 2). Cette complexité n'est pas étonnante étant donné les non-linéarités existant dans le lien entre l'articulatoire et l'acoustique. Cela suggère qu'une identification de $P(S_p^\Phi \mid \Phi)$ au résultat exact de l'inférence de l'Équation (2) demande une caractérisation proprioceptive complexe des phonèmes. Nous faisons l'hypothèse que cette complexité n'est pas stockée dans sa totalité dans la caractérisation des buts proprioceptifs associés à chaque phonème, et nous choisissons de définir $P(S_p^\Phi \mid \Phi)$ à partir d'une approximation par une mixture de quatre gaussiennes des échantillons de la Figure 2, dont on ne conserve au final que le mode gaussien de moindre variance. Les lignes verticales en pointillés sur la Figure 2 indiquent la position du noyaux gaussien retenu.

3.2 Comparaison des simulations des differentes inférences

Les trois inférences du modèle fournissent des distributions caractérisant les valeurs des commandes motrices (variable M) permettant de produire chaque phonème. Nous décrirons et comparerons dans cet article uniquement les résultats obtenus par les planifications dans l'espace auditif seul et dans l'espace proprioceptif seul. Ces distributions ont été approximées par un échantillonnage de type MCMC dans l'espace des commandes motrices. Pour comparer les deux stratégies de planification, nous étudions les résultats auditifs et proprioceptifs correspondants aux commandes motrices obtenues par chacune des inférences.

On s'attend à ce que la planification selon la modalité auditive donne lieu à une plus grande variabilité des configurations articulatoires, étant donné que la complexité des régions proprioceptives associées aux régions cibles dans l'espace acoustique nous a amenés à ne sélectionner qu'une zone restreinte dans chacune d'elles pour spécifier sous forme gaussienne les cibles à atteindre lors d'une planification

proprioceptive. Les images de droite de la Figure 3 présentent les résultats proprioceptifs [2] résultant de 100 échantillons de commandes motrices planifiées [3], selon la modalité proprioceptive (bas) et la modalité auditive (haut). Les ellipses grises représentent les ellipses de dispersion des données proprioceptives obtenues. Les ellipses noires représentent les cibles gaussiennes retenues pour la définition des cibles proprioceptives pour chaque phonème. Conformément à nos attentes, on constate que la planification dans l'espace proprioceptif donne des résultats dans cet espace conformes à la cible planifiée, alors que la planification dans l'espace auditif donne lieu à des configurations articulatoires nettement plus variables.

Dans le cadre de notre modélisation, la planification proprioceptive donne donc lieu à des formes du conduit vocal moins variables et plus prototypiques que la planification auditive. Du fait de la réduction de la région cible dans l'espace proprioceptif, que nous avons été amenés à proposer à cause de la complexité géométrique de la projection dans cet espace des régions cibles auditives, la planification proprioceptive n'autorise pas le recours à toutes les possibilités d'équivalence motrice qu'offre la planification auditive, à la fois dans le domaine des activations musculaires et dans le domaine des formes de langue. Dans ce contexte, la question qui se pose est celle de savoir si les deux modes de planification se différencient aussi par des variabilités différentes dans le domaine auditif. La réponse est donnée par les images de gauche de la Figure 3. Celles-ci présentent les résultats auditifs obtenus à partir des mêmes échantillons de commandes motrices utilisées pour générer les résultats proprioceptifs illustrés sur les images de droite. A nouveau, les ellipses grises représentent les ellipses de dispersion des données auditives obtenues. Les ellipses noires représentent les distributions gaussiennes caractérisant les cibles auditives pour chaque phonème. Nous constatons, et ce n'est pas une surprise, que les résultats obtenus par planification auditive sont en cohérence avec la cible planifiée. En revanche, nous observons de plus que les résultats obtenus par planification proprioceptive présentent globalement une variabilité similaire à celle obtenue par planification auditive [4], alors même que les formes de langue sont moins variables en planification proprioceptive. Cela indique que, même si l'ensemble des possibilités d'équivalence motrice n'est pas exploité par la planification proprioceptive, la variabilité des formes articulatoires autorisée par cette planification permet de générer une variabilité acoustique similaire à celle de la planification auditive. La plus grande variabilité proprioceptive liée à la planification auditive suggère une meilleure exploitation des possibilités d'équivalence motrice.

4 Discussion et conclusion

La comparaison des planifications vers des cibles gaussiennes définies dans l'espace proprioceptif et dans l'espace auditif a permis de mettre en évidence deux points importants. Tout d'abord l'approximation par un noyau gaussien des régions proprioceptives associées aux cibles auditives, elles

[2] Le sous espace des deux premières composantes principales des longueurs de muscles, obtenues sur l'ensemble des réalisations des phonèmes, est choisi pour visualisation. Cette caractérisation est pertinente étant donné que la dispersion des données dans l'espace des 6 longueurs de muscles est expliquée à plus de 90% par les trois premières composantes principales. On reconnait d'ailleurs dans les deux premières composantes extraites les dimensions haut/bas et avant/arrière à partir desquelles on positionne les voyelles dans l'espace du conduit vocal (Harshman *et al.*, 1977).

[3] Pour des raisons de clarté de la représentation, seuls 100 échantillons tirés aléatoirement parmi l'ensemble des commandes obtenues par l'échantillonnage MCMC ont été représentées.

[4] Les résultats auditifs obtenus pour le phonème /ɛ/ en planification proprioceptive présentent une variabilité bien plus importante où on retrouve une partie importante des résultats auditifs qui s'écartent de la cible correspondante. Ceci met en évidence des discontinuités ou non convexités de la projection proprioceptive des cibles auditives que le noyau gaussien utilisé retient à tort. Ce phénomène a vraisemblablement pour origine l'imprécision des modèles internes utilisés.

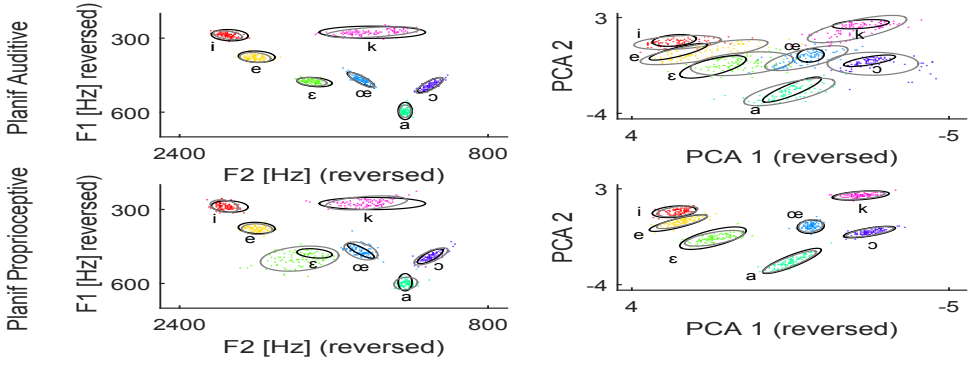


FIGURE 3 : Réalisations auditives (gauche) et proprioceptives (droite) obtenues par le modèle biomécanique de langue à partir de 100 commandes motrices obtenues par planification auditive (haut) et planification proprioceptive (bas) pour les sept phonèmes considérés. Les ellipses grises représentent les ellipses de dispersion des données générées. Les ellipses noires représentent les noyaux gaussiens caractérisant les cibles auditive (images de gauche) et proprioceptive (images de droite) pour chaque phonème

aussi gaussiennes, ne permet pas d'extraire toute la complexité de la distribution des configurations articulatoires donnant lieu à un même phonème. La planification dans l'espace proprioceptif donne donc naturellement des réalisations d'un même phonème moins variables, plus prototypiques. Cette planification ne permet pas l'exploitation de toutes les possibilités d'équivalence motrice.

On peut alors s'attendre à ce que la planification de séquences de phonèmes dans le domaine proprioceptif révèle une coarticulation de moins grande amplitude qu'une planification dans l'espace auditif. Ensuite, malgré la restriction des configurations articulatoires imposée par notre hypothèse de planification proprioceptive, la variabilité auditive observée est qualitativement similaire à celle générée par la planification auditive, bien que cette dernière exploite davantage de configurations articulatoires. Ainsi, d'un point de vue purement acoustique nos résultats suggèrent que, pour ce qui est de la production de phonèmes isolés du moins, les deux stratégies de planification seraient difficiles à distinguer. Dans la production de séquences de phonèmes cependant, on s'attend à ce qu'une planification reposant sur des cibles proprioceptives prototypiques induise des patrons de coarticulation différents.

Globalement, la modélisation que nous proposons des cibles auditives ou proprioceptives, ainsi que de la planification auditive ou proprioceptive, offre un cadre précis qui permet de prédire les propriétés des phonèmes planifiés selon l'une ou l'autre de ces modalités. Il convient maintenant d'évaluer la validité de ces prédictions par des expériences où des sujets seront placés dans des conditions favorisant l'une ou l'autre de ces planifications. C'est ce à quoi nous nous attachons désormais.

Remerciements

Ces recherches ont bénéficié du soutien financier 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).

Références

BADIN P. & FANT G. (1984). Notes on vocal tract computation. *Quarterly Progress and Status Report, Dept for Speech, Music and Hearing, KTH, Stockholm*, p. 53–108.

BESSIÈRE P., MAZER E., AHUACTZIN J. M. & MEKHNACHA K. (2013). *Bayesian Programming*. Boca Raton, Florida : CRC Press.

BLAKEMORE S.-J., WOLPERT D. & FRITH C. (2000). Why can't you tickle yourself ? *Neuroreport*, **11**(11), R11–R16.

BROWMAN C. P. & GOLDSTEIN L. (1992). Articulatory phonology : An overview. *Phonetica*, **49**(3-4), 155–180.

CALLAN D. E., JONES J. A., CALLAN A. M. & AKAHANE-YAMADA R. (2004). Phonetic perceptual identification by native-and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory–auditory/orosensory internal models. *NeuroImage*, **22**(3), 1182–1194.

FELDMAN A. G. (1986). Once more on the equilibrium-point hypothesis (λ model) for motor control. *Journal of motor behavior*, **18**(1), 17–54.

GUENTHER F. H., GHOSH S. S. & TOURVILLE J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and language*, **96**(3), 280–301.

HARSHMAN R., LADEFOGED P. & GOLDSTEIN L. (1977). Factor analysis of tongue shapes. *The Journal of the Acoustical Society of America*, **62**(3), 693–707.

HICKOK G. (2014). The architecture of speech production and the role of the phoneme in speech processing. *Language, Cognition and Neuroscience*, **29**(1), 2–20.

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

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.

PALIWAL K. K., AINSWORTH W. A. & LINDSAY D. (1983). A study of two-formant models for vowel identification. *Speech Communication*, **2**(4), 295–303.

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.

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. *The Journal of the Acoustical Society of America*, **114**(3), 1582–1599.

SALTZMAN E. L. & MUNHALL K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological psychology*, **1**(4), 333–382.

STEVENS K. N. (1993). Models for the production and acoustics of stop consonants. *Speech communication*, **13**(3), 367–375.

Une méthode d'évaluation de la compréhension orale par choix d'image : application à de la parole dégradée par simulation de la presbyacousie

Cynthia Magnen[1] Julien Tardieu[1] Lionel Fontan[2] Pascal Gaillard[3] Nathalie Spanghero-Gaillard[4]
(1) MSHS-T USR3414, Université de Toulouse et CNRS, Toulouse, France
(2) ARCHEAN Technologies, Montauban, France
(3) CLLE UMR5263, Université de Toulouse et CNRS, Toulouse, France
(4) LERASS-CPST EA827, Université de Toulouse, Toulouse, France
cynthia.magnen@univ-tlse2.fr, julien.tardieu@univ-tlse2.fr, lfontan@archean.fr, pascal.gaillard@univ-tlse2.fr, nathalie.spanghero@univ-tlse2.fr

RESUME

Nous présentons une méthode permettant d'évaluer la compréhension de la parole dégradée par simulation des effets de la presbyacousie, dans le calme et dans le bruit. Cette méthode intègre des phrases signifiantes et implique pour l'auditeur de sélectionner, parmi un ensemble de quatre images, celle qui correspond à l'énoncé qu'il entend. Le test présente de nombreux avantages méthodologiques comme l'immédiateté du score et le fait qu'il ne nécessite pas de faire répéter la phrase entendue. Les résultats obtenus montrent un effet significatif de la dégradation et du bruit du fond. La cohérence de ces effets avec les études précédentes sur la presbyacousie permet de valider cette méthode. Par ailleurs, la nature exacte du score mesuré dans ce test est discutée en le comparant avec le score d'intelligibilité obtenu par répétition d'items dans une précédente étude.

ABSTRACT

A method for assessing listening comprehension using image selection : application to speech degraded by presbycusis simulation.
This paper presents a method for assessing the comprehension of speech degraded by presbycusis simulation in both quiet and babble noise. This method uses meaningful sentences and asks the participant to select, among four images, the one that best corresponds to what he/she hears. This test has many methodological advantages such as : immediate score and the participant is not asked to repeat the sentence. The results show a significant effect of both the degradation and the background noise. These effects are coherent with previous studies on presbycusis which provides a validation of our method. In addition, the exact nature of the score measured in this test is discussed by comparing it to the intelligibility score measured with repeated sentences in a previous study.

MOTS-CLES : Presbyacousie, intelligibilité, compréhension, perception, parole dans le bruit.
KEYWORDS: Presbycusis, intelligibility, comprehension, perception, speech in noise.

1 Introduction

On pense souvent, et assez légitimement, qu'une bonne audition est garante de la compréhension orale d'une situation de communication. De sorte que si l'on veut étudier les processus sous-jacents à la compréhension, il semble évident de décrire en premier lieu les mécanismes liés à l'audition. Dans l'étude des surdités, par exemple, afin de mesurer les pertes auditives des patients et le gain apporté par l'appareillage auditif, les audioprothésistes évaluent l'intelligibilité de la parole en recourant à des tests de perception de mots ou de phrases. Or, plusieurs travaux ont contribué à montrer que la perception auditive ne peut être vue comme une simple activité d'identification. Les études sur les phénomènes de restauration phonémique (Sivonen *et al*., 2006) et d'illusions sémantiques (Nieuwland, Van Berkum, 2005) ont notamment traduit l'existence d'attentes perceptives chez l'auditeur en montrant que des processus « descendants » (*top-down*, du sens vers la forme) influencent les traitements dits « de plus bas niveau ». C'est-à-dire qu'une information de nature auditive est identifiable par un auditeur si elle intégrée dans la dynamique d'une situation et donc, déjà saisie dans un projet d'utilisation.

1.1 La mesure de l'intelligibilité de la parole en audiométrie vocale

Dans le domaine de la réhabilitation auditive, les capacités de discrimination phonémique et de compréhension du langage de patients atteints de troubles auditifs sont mesurées grâce à un examen d'audiométrie vocale. Cet examen est complémentaire de l'audiométrie tonale qui détecte les seuils absolus de perception de sons purs. En plus de s'intéresser aux répercussions sociales de la perte auditive, l'audiométrie vocale permet de différencier entre les distorsions d'origine endocochléaire et les atteintes centrales de l'audition. Le diagnostic ainsi établi permet de choisir une technique de réhabilitation adaptée (Bouccara *et al*. 2005). En seconde intention, le praticien juge l'efficacité des moyens de compensation mis en place tels que la prothèse ou l'implant cochléaire par un nouvel examen (Bonfils, Avan, 2005).

Les tests utilisés dans cette optique permettent d'établir un score d'intelligibilité en cotant en pourcentage le taux de reconnaissance du matériel verbal présenté au sein d'une liste (mots, phrases, logatomes). Selon les pratiques, les listes sont soit enregistrées, soit énoncées par l'expérimentateur. Au cours de ces tests, on demande aux patients de répéter des listes de stimuli sonores émis à différentes intensités par voie aérienne ou par voie osseuse. Ainsi, sont décrits : 1) le seuil d'intelligibilité, intensité à partir de laquelle 50 % des mots sont compris, 2) le maximum d'intelligibilité, pourcentage maximum de mots compris et 3) le pourcentage de discrimination, pourcentage de stimuli sonores compris à 35 dB au dessus du seuil d'intelligibilité. L'audiométrie vocale peut aussi se réaliser avec l'adjonction de bruit perturbant afin de confirmer les scores obtenus pour les tests en cabine (les scores étant souvent inférieurs dans le bruit).

1.2 La diversité du matériel linguistique

Le matériel sonore utilisé est très variable. Il peut s'agir de listes de sons complexes le plus souvent signifiants et possiblement non signifiants. Les plus utilisées en France sont les listes de mots dissyllabiques de Fournier (1951) et les listes de mots de 3 phonèmes de Lafon (1964). En proposant la répétition erronée du mot comme unité d'erreur, les listes de Fournier ont pour objectif de mesurer l'intelligibilité. Les listes de Lafon mesurent quant à elles l'identification phonémique par un comptage du nombre de phonèmes perçus ou erronés. De nombreuses critiques

ont été formulées sur l'élaboration de ces listes. Parmi elles, nous pouvons évoquer la désuétude du vocabulaire, le français ayant beaucoup évolué en une cinquantaine d'années, la fréquence d'occurrence des mots employés s'en trouve modifiée (Garnier *et al*., 1997). L'ordonnancement interne des items a également été soulevé puisque la présentation successive de différents termes peut induire des effets d'amorçage sémantique (Estienne, Piérart, 2006). Enfin, la qualité de l'équilibrage phonétique dans les listes de Fournier (Lafon, 1964), comme l'inégalité de longueur des stimuli pour les listes de Lafon et le manque de reproductibilité des tests (Garnier *et al*., 1997) sont autant de limites évoquées dans l'utilisation clinique de ce matériel.

Jusqu'à il y a peu, le contexte phrastique était moins utilisé que le mot pour évaluer l'intelligibilité de la parole. Paradoxalement, les effets de la suppléance mentale étaient considérés comme un inconvénient dans l'évaluation des capacités auditives réelles. D'autant que plus l'unité traitée est grande plus la suppléance mentale est influente dans la réalisation de la tâche (Garnier *et al*. 1997). Aujourd'hui, l'utilisation de la parole dans ce contexte gagne en popularité, car elle est plus représentative des situations de communication quotidiennes (indices contextuels et bruit de fond) que les listes de mots isolés diffusées en milieu calme. Les tests de répétition de phrases semblent donc plus valides du point de vue de la prédictibilité (Schoepflin, 2015). Le plus populaire est le test HINT (pour *Hearing in Noise Test* - Nilsson *et al*., 1994) qui évalue la reconnaissance de 250 phrases (25 listes de 10 phrases) variant de 3 à 7 mots; le score est calculé à partir du nombre total de mots correctement identifiés dans les phrases.

1.3 Objectif de l'étude

L'objectif de l'étude est de présenter une nouvelle méthode permettant d'évaluer la compréhension de la parole dégradée par simulation des effets de la presbyacousie, dans le calme et dans le bruit. Cette méthode intègre des phrases signifiantes et propose aux participants d'apparier ces phrases à des images qui les illustrent. Afin de valider notre méthode, nous comparerons les scores de compréhension par choix d'images aux scores obtenus dans une étude précédente avec une tâche de répétition de phrases (*cf*. Fontan *et al*., 2014). L'étude montrait notamment une baisse significative du score de reconnaissance des phrases en fonction de la dégradation et du bruit de fond.

2 Le test de compréhension de la parole

2.1 Choix d'un test d'appariement phrase/image

Afin d'évaluer la compréhension de la parole dégradée, nous avons choisi d'élaborer un test d'appariement énoncé oral / image. Ce type de test est déjà utilisé dans une visée diagnostique pour évaluer la capacité de patients aphasiques à traiter des énoncés dont la complexité syntaxique va croissant (*cf.* Paradis, 1989), ainsi qu'en didactique des langues pour mesurer le niveau de compréhension d'un apprenant de langue seconde (Spanghero-Gaillard, 2008). La tâche proposée implique pour l'auditeur de sélectionner, parmi un ensemble de quatre images, celle qui correspond à l'énoncé qu'il entend. Elle présente l'avantage de proposer un contexte commun au locuteur et à l'auditeur représenté par un ensemble visuel de référents possibles (le jeu d'images). Le locuteur y fait référence dans son énoncé, référence à laquelle doit répondre l'auditeur par une tâche de sélection. En adaptant ce test à la compréhension de la parole dégradée, nous ne testons

plus la capacité à répéter des mots ou des phrases. Nous nous intéressons à la compréhension d'un énoncé dégradé dans une situation de communication s'approchant davantage de la vie réelle, c'est-à-dire où le contexte n'est pas figé mais sujet à interprétation. Ce cadre est avantageux mais il n'est pas évident à mettre en place car il exige de prédéfinir un ensemble de réponses limité. Les critères retenus pour constituer le matériel expérimental sont présentés ci-après.

2.2 Élaboration du matériel linguistique

L'élaboration des variantes de l'énoncé n'est pas une tâche triviale. Il s'agit de définir des ambiguïtés reflétant la perception de l'énoncé dégradé par simulation des effets de la presbyacousie. Pour résoudre cette difficulté, nous avons utilisé les données d'un test précédent (*cf.* Test 2 dans Fontan *et al.,* 2014). Le test consistait à faire répéter des phrases issues de la version française du test HINT (Vaillancourt *et al*., 2005) à une trentaine d'auditeurs francophones (âgés entre 18 et 30 ans). Les phrases étaient dégradées selon les mêmes conditions que pour cette présente étude (*cf.* section 3.3). Les résultats constituent un corpus d'énoncés soit conformes d'un point de vue phonétique et sémantique, soit déviants par rapport à l'énoncé de base. Au sein de ce corpus, nous avons sélectionné les 33 phrases qui pouvaient être facilement illustrées (ex.: « L'oiseau s'envole du nid »). Puis, pour chacune des phrases sélectionnées, nous avons relevé les trois énoncés déviants les plus fréquemment cités (ex. : « L'oiseau s'envole de nuit », « Les oiseaux s'envolent du nid », « Les oiseaux sont près du nid »). En fonction du niveau de dégradation entendu, les énoncés produits pouvaient donner lieu à des scènes irréelles plutôt rigolotes et incongrues (ex. pour « Les dragons crachent du feu »: « Le dragon crache du feu », « Les dragons crachent des pneus », « Le chaton crache du feu »). En tout, le corpus audio est constitué de 132 phrases assertives (3 variantes de 33 énoncés, pour 12 phrases d'entraînement et 120 phrases de test).

2.3 Représentations imagées des énoncés : conception, réalisation et accord

Les phases de conception et de réalisation des représentations imagées des 132 énoncés ont été menées en étroite collaboration avec deux graphistes professionnels (Jérémy Villy[1] et Olivier Subra[2]). Des croquis réalisés par nos soins leur ont été soumis en première intention afin de préciser le cahier des charges. D'un point de vue de la réalisation, il était important que le participant puisse interpréter le sens du dessin sans qu'il n'y ait d'ambiguïté sur la signification graphique. Les illustrateurs ont donc choisi de travailler en ligne claire, avec des aplats de couleur en reprenant les codes historiques de la bande dessinée, popularisés notamment avec Tintin, Lucky Luke ou Spirou. Un pré-test a été mené auprès de cinq personnes francophones natives âgées entre 20 et 30 ans. Ce pré-test a consisté à apparier chacune des phrases écrites aux 4 possibilités visuelles. Chaque phrase était présentée en même temps que les 4 images correspondantes sur un écran informatique via une interface dédiée. Une fois l'image choisie, les participants évaluaient leur réponse sur une échelle indicielle de confiance allant de 1 à 7. Lorsque l'indice de confiance était en deçà de 5, le participant justifiait son choix par un commentaire libre écrit. Le matériel visuel a été modifié en fonction des résultats de ce test et grâce au recueil des commentaires afin qu'aucune ambiguïté ne subsiste quant à l'appariement phrase/image.

[1] http://jeremyvilly.com/

[2] http://www.admarginem.fr/

3 Méthode

3.1 Participants

Trente participants ont été sélectionnés et rémunérés pour cette expérience. Le profil des participants répondait aux critères d'inclusion suivants : étudiants francophones natifs, âgés de 18 à 30 ans inclus, sans problème de vue non corrigé par des lentilles ou des lunettes. Le niveau d'audition de chaque participant a été vérifié par un audiogramme tonal (critère d'inclusion : moyenne des pertes entre 2kHz et 8kHz < 15 dB).

3.2 Stimuli

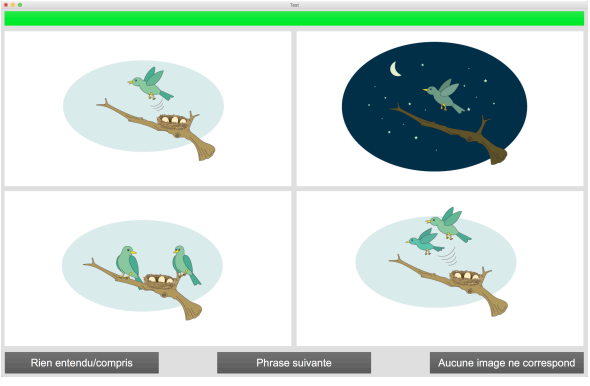


FIGURE 1. Exemple de fenêtre présentant quatre images ainsi que les boutons de réponses pour le test de compréhension de phrases.

Audio. L'ensemble des 132 phrases (*cf*. section 2.2 pour la description du matériel linguistique) a été enregistré auprès de deux locuteurs adultes (un homme et une femme de 46 et 47 ans) et d'une enfant (12 ans). L'enregistrement a eu lieu dans la cabine audiométrique PETRA (http://petra.univ-tlse2.fr), avec un microphone omnidirectionnel Sennheiser MD46, une console de mixage TASCAM DM-3200 et un ordinateur MacPro équipé du logiciel Reaper. Les phrases d'une durée moyenne de 1273 ms (écart type = 186 ms) ont ensuite été égalisées en sonie par un panel d'auditeurs puis mixées avec un bruit de fond vocal de type « brouhaha » et un rapport signal sur bruit de 5 dBA. Les phrases avec et sans bruit de fond ont enfin été dégradées par simulation de la presbyacousie suivant une variation en dix niveaux. Le premier niveau correspond au signal non dégradé et les 9 niveaux suivant vont d'un âge théorique de 60 ans à 110 ans. La procédure de traitement de signal est détaillée dans Fontan *et al*. (2014).

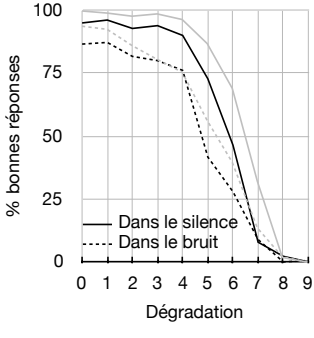
Visuels. Les images décrites dans la section 2.3 sont présentées par planches de quatre images de 820x440 pixels sur un écran Apple Cinema Display de 20 pouces avec une résolution de 1680x1050 pixels. La fenêtre présentant les quatre images et les boutons de réponse occupe tout l'écran et est illustrée en Figure 1.

3.3 Procédure

Le test de compréhension a été effectué dans la cabine audiométrique PETRA (*cf.* section 3.3). Les participants étaient assis à un mètre des haut-parleurs (Tannoy Precision 6D), avec un niveau de diffusion des phrases non dégradées (condition 0) de 60 dBA et un niveau de bruit de fond vocal pour les phrases non dégradées à 55 dBA. L'interface a été créée avec le logiciel Max/MSP permettant la présentation concomitante des stimuli visuels et audio (voir Figure 1). Après un entrainement sur 12 phrases, les participants entendaient les 120 phrases de test. Ils avaient pour consigne de sélectionner l'image correspondant à la phrase entendue. Ils étaient également averti que les énoncés qu'ils entendraient pouvaient être totalement incongrus. Une barre latérale verte s'allumait en haut de l'écran pour indiquer la diffusion de la phrase. Dès lors que la barre latérale était active, ils devaient répondre le plus rapidement possible, avec la possibilité de répondre avant la fin de la diffusion de la phrase (les temps de réaction sont mesurés à partir du début de chaque phrase). Un bouton « Phrase suivante » permettait de passer à une nouvelle phrase. Si les participants jugeaient que le signal de parole n'était pas compréhensible ou audible, ils pouvaient sélectionner le bouton « rien entendu/compris ». Si aucune image ne correspondaient à ce qu'ils avaient perçu, les participants pouvaient sélectionner le bouton « Aucune image ne correspond ». L'enregistrement des temps de réaction et des réponses choisies par les participants a été effectué via le logiciel de passation dédié. Le plan expérimental comporte 2 facteurs intra sujet : la dégradation (10 niveaux) et la condition de bruit de fond (2 niveaux).

4 Résultats

4.1 Scores de compréhension



	% bonnes réponses	
Dég.	Silence	Bruit
0	95,0 (21,8)	86,6 (33,4)
1	96,1 (19,4)	87,2 (31,0)
2	92,8 (26,1)	81,7 (38,9)
3	93,9 (23,4)	80,0 (40,2)
4	90,0 (29,9)	76,1 (42,9)
5	72,8 (44,7)	41,7 (49,7)
6	46,4 (49,9)	27,9 (44,8)
7	7,8 (25,1)	8,9 (27,7)
8	2,2 (14,7)	0,0 (0,0)
9	0,0 (0,0)	0,6 (4,3)

FIGURE 2. Pourcentages de bonnes réponses en fonction du niveau de dégradation, dans le silence et dans le bruit, écarts types entre parenthèses. En grisé, sont présentés les résultats du test d'intelligibilité par répétition de phrases issus de Fontan *et al*. (2014).

Le score est de 100 % si l'appariement énoncé/image est correctement réalisé ou 0 % sinon (même si l'image et le stimulus audio présentent des éléments communs). Les résultats sont présentés sur la Figure 2 et montrent que plus la dégradation du signal de parole augmente, plus les scores de compréhension diminuent, dans le silence comme dans le bruit (différence moyenne = 10,8 %). Les données ont été analysées en utilisant un modèle linéaire mixte généralisé et ont révélé un effet significatif de la dégradation (χ^2 (9) = 2416,7; p < 0,001) et du bruit (χ^2 (1) = 95,3; p < 0,001). La Figure 2 présente en grisé les scores obtenus dans (Fontan *et al*. 2014) avec une tâche

de répétition de phrases issues du test HINT, la comparaison entre ces deux résultats est discutée en section 5.

4.2 Type de réponse et temps de réaction

L'analyse du type de réponses permet d'observer le nombre de fois que les participants ont sélectionné un des trois types de réponses possibles: choix d'images, bouton « Aucune image ne correspond » et bouton « Rien entendu/compris ». Les résultats montrent que les participants n'ont quasiment jamais utilisé le bouton « Aucune image ne correspond » (1 fois dans le silence et 6 fois dans le bruit sur une totalité de 3600 réponses). La Figure 3 (Gauche) présente le nombre de fois (en %) que les participants ont utilisé le bouton « Rien entendu/compris » en fonction de la dégradation et du bruit de fond. Ce pourcentage dépasse les 50% à partir des dégradations 6 et 7, limite au delà de laquelle les participants n'arrivent plus à choisir une des images proposées.

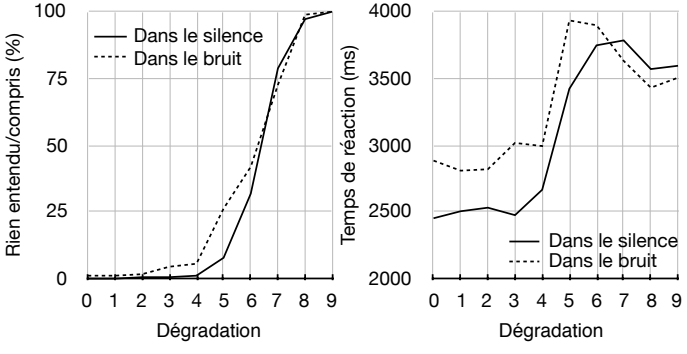


FIGURE 3. Gauche : nombre d'appui sur le bouton « Rien entendu/compris » (en %) en fonction de la dégradation. Droite : Temps de réaction (ms) en fonction de la dégradation. Toutes les réponses de participants sont prises en comptes.

Dans l'analyse des temps de réaction, toutes les réponses des participants sont prises en compte (Figure 3 - Droite) afin d'observer l'évolution de la difficulté de la tâche dans son ensemble. Les données ont été analysées à l'aide d'un modèle linéaire mixte et montrent un effet significatif de la dégradation ($F(9; 3504) = 53{,}4$; $p < 0{,}001$) et du bruit de fond ($F(1; 3504) = 141{,}7$; $p < 0{,}001$). Les données dans le silence présentent une forte augmentation du temps de réaction à partir de la dégradation 4 puis une baisse à partir de la dégradation 7. Le bruit de fond est aussi responsable d'une forte augmentation des temps de réaction avec une augmentation moyenne de 217 ms.

5 Discussion

Le test de compréhension proposé dans ce papier diffère des tâches habituellement proposées dans les tests d'intelligibilité en audiométrie vocale car il s'agit pour le participant de sélectionner une image qui correspond à l'énoncé entendu. Les résultats obtenus avec de la parole dégradée par simulation de la presbyacousie montrent un effet significatif de la dégradation et du bruit du fond. Ces effets sont cohérents avec les études précédentes sur les effets de la presbyacousie sur l'audition (Moore, 2007, Fontan *et al*. 2014), c'est-à-dire une baisse des scores en fonction de la gravité de la presbyacousie et de la présence de bruit de fond vocal. Par ailleurs, la comparaison

des scores de compréhension obtenus dans ce test avec les scores d'intelligibilité obtenus par répétition de phrases dans Fontan et al. (2014) nous permet de discuter la nature exacte du score mesuré. D'une part, les scores de compréhension sont plus faibles que les scores d'intelligibilité, avec des différences jusqu'à 24% à la dégradation 7 dans le silence (*cf.* figure 2). Ainsi, les participants sont meilleurs en répétition de phrases qu'en choix d'images. D'autre part, les temps de réaction observés dans les deux études présentent une augmentation puis un baisse entre les dégradations 4 et 7 mais avec une augmentation totale de 1119 ms dans le test de compréhension contre 321 ms dans le test d'intelligibilité.

Les différences entre les deux tâches traduisent la complexité de la tâche de choix d'images. Pour y répondre, le participant doit accomplir trois objectifs en parallèle et de façon dynamique: interpréter un message acoustique plus ou moins dégradé, interpréter des scènes visuelles et enfin apparier l'énoncé compris à l'image correspondante. Il semblerait que deux stratégies soient mises en place par les participants pour répondre à ces objectifs. Lorsque la dégradation est faible (en dessous de 4), les participants entendent le message et l'interprète facilement. L'appariement entre l'énoncé et l'image correspondante est donc très rapide car le participant sélectionne simplement les éléments visuels correspondant à ce qu'il a entendu. À partir d'une dégradation plus importante (à partir de 4), la tâche devient plus difficile car le participant cherche des indices lui permettant de restaurer les éléments phonétiques manquants dans le message acoustique. Ces indices sont disponibles parmi les quatre scènes imagées qui présentent une grande diversité d'éléments visuels. Cette complexité est accrue par le fait que les scènes visuelles sont très proches sémantiquement. Des mesures occulométriques permettraient d'étudier plus en détail ces deux stratégies en observant les zones et durées de focalisation ou la dilatation pupillaire par exemple.

6 Conclusions et perspectives

Le test de compréhension proposé dans cette étude permet d'obtenir des scores cohérents avec les résultats obtenus par des tests plus classiques (répétition de mots ou de phrases) dans le domaine de l'audiométrie. Il se présente comme un test plus facile à mettre en oeuvre (pas besoin de faire répéter, pas besoin de transcrire les réponses, adapté à un public plus vaste) tout en se rapprochant d'une situation de communication réelle par l'utilisation de phrases et d'un contexte visuel. La complexité de la tâche (faire correspondre ce qui est entendu avec les indices trouvés dans les images) illustre une partie des difficultés rencontrées par les patients atteints de presbyacousie dans leur vie quotidienne. Les scores plus faibles en compréhension que dans les tests d'intelligibilité plus classiques permettent d'interroger le lien entre intelligibilité et compréhension. Dans notre cas, l'intelligibilité surestime la compréhension c'est-à-dire qu'au-delà d'une certaine dégradation, les scores d'intelligibilité ne permettent pas de savoir si le message a bien été compris. Ce résultat est à explorer davantage au regard des résultats contradictoires obtenus dans des études précédentes interrogeant également ce lien (*cf.* Fontan *et al.*, 2015).

Remerciements

Cette étude a été réalisée dans le cadre d'un projet financé par la Région Midi-Pyrénées.

Références

BONFILS P. et AVAN P. (2005). Evaluation du système auditif. In Dulguerov P. & Remacle M. *Précis d'audiophonologie et de déglutition, Tome 1 : l'oreille et les voies de l'audition* (pp.149-163). Marseille : Solal.

BOUCCARA D., AVAN P., MOSNIER I., BOZORG GRAYELI A., FERRARY E. et STERKERS O. (2005). Réhabilitation auditive, *Medecine Sciences* 21(2), 190-197.

ESTIENNE F. et PIÉRART B. (2006). *Les bilans de langage et de voix. Fondements théoriques et pratiques*. Paris : Masson.

FONTAN L., MAGNEN C., TARDIEU J., et GAILLARD P. (2014). Simulation des effets de la presbyacousie sur l'intelligibilité et la compréhension de la parole dans le silence et dans le bruit. *Actes des Journées d'étude sur la parole (30e édition, Le Mans)*, 694-702.

FONTAN L., TARDIEU J., GAILLARD P. et RUIZ R. (2015). Relationship Between Speech Intelligibility and Speech Comprehension in Babble Noise. *Journal of Speech Language and Hearing Research* 58, 977-986.

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*. Maloine.

GARNIER T., VESSON J.F., AUDRY J.C. et AZEMA B. (1997). Épreuves vocales, applications. In *Précis d'audioprothèse: l'appareillage de l'adulte. Tome 1, Le bilan d'orientation prothétique* (pp. 203-247). Paris : Éditions du Collège National d'Audioprothèse (1ère éd. 1997).

LAFON J.C. (1964). *Le test phonétique et la mesure de l'audition*. Eindhoven : Éditions Centrex.

MOORE B. C. J. (2007). *Cochlear hearing loss : Physiological, psychological and technical Issues*. Wiley.

NILSSON M., SOLI S. et SULLIVAN J.A. (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America* 95(2), 1085-1099.

NIEUWLAND M. S. et VAN BERKUM J. J. A. (2005). Testing the limits of the semantic illusion phenomenon: ERPs reveal temporary semantic change deafness in discourse comprehension. *Cognitive Brain Research* 24(3), 691-701.

PARADIS M. (1989). *Bilingual Aphasia test* (French version). Mahwah : Lawrence Erlbaum.

SCHOEPFLIN J. (2015). Back to basics : Speech Audiometry, *http://www.audiologyonline.com/*.

SIVONEN P., MAESS B., LATTNER S. et FRIEDERICI A. D. (2006). Phonemic restoration in a sentence context: Evidence from early and late ERP effects. *Brain Research* 1121(1), 177-189.

SPANGHERO-GAILLARD N. (2008). Comment l'être humain comprend? Quelques éléments de réflexion à partir de l'apprentissage d'une langue étrangère, Habilitation à Diriger des Recherches, Spécialité Sciences du Langage, Université Toulouse - Le Mirail, Toulouse.

VAILLANCOURT V., LAROCHE C., MAYER C., BASQUE C., NALI M., ERIKS-BROPHY A., SOLI S. D. et GIGUERE C. (2005). Adaptation of the HINT (Hearing in Noise Test) for adult Canadian Francophone populations. *International Journal of Audiology* (44), 358-369.

Optimiser l'adaptation en ligne d'un module de compréhension de la parole avec un algorithme de bandit contre un adversaire

Emmanuel Ferreira, Alexandre Reiffers-Masson, Bassam Jabaian et Fabrice Lefèvre
CERI-LIA, Université d'Avignon et des Pays de Vaucluse, France
prenom.nom@univ-avignon.fr

RÉSUMÉ

De nombreux modules de compréhension de la parole ont en commun d'être probabilistes et basés sur des algorithmes d'apprentissage automatique. Deux difficultés majeures, rencontrées par toutes les méthodes existantes sont : le coût de la collecte des données et l'adaptation d'un module existant à un nouveau domaine. Dans cet article, nous proposons un processus d'adaptation en ligne avec une politique apprise en utilisant un algorithme de type bandit contre un adversaire. Nous montrons que cette proposition peut permettre d'optimiser un équilibre entre le coût de la collecte des retours demandés aux utilisateurs et la performance globale de la compréhension du langage parlé après sa mise à jour.

ABSTRACT

Adversarial bandit for optimising online active learning of spoken language understanding

Many speech understanding modules have in common to be probabilistic and to rely on machine learning algorithms to train their models from large amount of data. The difficulty remains in the cost of collecting such data and the time for updating an existing model to a new domain. In this paper, we propose to drive an online adaptive process with a policy learnt using the Adversarial Bandit algorithm. We showthat this proposition can optimally balance the cost of gathering valuable user feedbacks and the overall performance of the spoken language understanding module after its update.

MOTS-CLÉS : Compréhension de la parole, Apprentissage sans données de références, Bandit contre un adversaire, Adaptation en ligne.

KEYWORDS: Spoken language understanding, zero-shot learning, Adversarial bandit problem, online adaptation.

1 Introduction

Dans un système de dialogue, le rôle du module de compréhension de la parole est d'extraire, pour chaque énoncé d'utilisateur, des hypothèses qui représentent son contenu sémantique. Cela peut être représenté par une séquence d'actes de dialogue sous la forme `acttype(slot=value)`. Les `acttype` sont indépendants de la tâche et transmettent l'intention de l'utilisateur lors de la communication, alors que les `slot` et les `value` dépendent du domaine d'application et correspondent à l'information que le système peut manipuler (entrées dans une base de données, commandes à un robot, ...). Par exemple, l'énoncé « Bonjour je cherche un restaurant français dans la partie sud de la ville » correspond à l'acte de dialogue « `hello(), inform(food=french), inform(area=south)` ».

Dans la dernière décennie, les systèmes de compréhension ont évolué progressivement vers des approches probabilistes apprises sur de grandes quantités de données (Hahn *et al.*, 2010; Deoras & Sarikaya, 2013). Plusieurs travaux ont proposé des approches semi-supervisées (Celikyilmaz *et al.*, 2011; Heck & Hakkani-Tur, 2012) ou non-supervisées (Tur *et al.*, 2011; Lorenzo *et al.*, 2013) pour faire face au manque de données annotées. Dans ce même but, d'autres travaux ont proposé de porter les systèmes à travers les langues et les domaines (Lefèvre *et al.*, 2012; Jabaian *et al.*, 2013). Plusieurs recherches ont aussi étudié des procédures d'apprentissage actif (Active Learning, AL) pour réduire le coût d'annotation et de vérification de corpus (Gotab *et al.*, 2010; Bayer & Riccardi, 2013).

Par ailleurs, dans (Ferreira *et al.*, 2015), nous avons proposé une méthode sans données de référence pour la compréhension de la parole. Cette méthode est basée sur une représentation vectorielle compacte de mots (word embedding (Mikolov *et al.*, 2013a)) utilisée pour généraliser une base de connaissance d'un domaine cible (bases de données, description ontologique...). Cette approche ne nécessite pas de données annotées et peut atteindre instantanément des performances état-de-l'art. Une stratégie d'adaptation en ligne a également été proposée pour affiner le modèle de façon progressive avec une supervision minimale. En effet, avec cette stratégie, les utilisateurs doivent confirmer certaines hypothèses faites par le système par des retours binaires, mais sans corriger explicitement les erreurs qui nécessitent des retours utilisateurs plus complexes. Cela permet de corriger certaines erreurs de classification mais sans possibilité d'ajouter de nouveaux concepts ou valeurs dans le modèle. Aussi, dans cet article, nous proposons d'étendre cette stratégie d'adaptation en ligne afin de répondre également à ce problème et d'être ainsi en mesure d'étendre le modèle avec de nouvelles connaissances en permanence.

Pour définir cette nouvelle stratégie, nous proposons de considérer le problème d'adaptation comme un problème de Bandit contre un Adversaire (Adversarial Bandit). Cette proposition vise à minimiser le coût de supervision tout en demandant à l'utilisateur des retours qui peuvent avoir le maximum d'impact sur le modèle. Les algorithmes de bandit ont été largement étudiés dans la communauté de l'apprentissage automatique (Auer *et al.*, 2002; Bubeck & Cesa-Bianchi, 2012). Leur objectif est de déterminer le meilleur compromis entre l'exploration des options qui ont donné le meilleur rendement (gains) dans les itérations précédentes et l'exploration de nouvelles options qui pourraient donner une meilleure performance à l'avenir. Peu de travaux ont déjà employé ce genre de techniques pour optimiser un module de traitement de langage naturel. Parmi ceux-ci, on pourra donner comme exemple (Ralaivola *et al.*, 2011) où les auteurs appliquent un algorithme de bandit contextuel pour raffiner un classificateur multiclasses avec des retours utilisateurs de type oui/non sur une tâche visant à identifier le motif d'un appel téléphonique (*call routing*). Nous montrons que la technique proposée dans ce papier permet d'obtenir de bonnes performances avec un coût faible et une supervision minime sur une tâche de compréhension de la parole en utilisant les données de la seconde campagne d'évaluation Dialog State Tracking Challenge (DSTC2) (Henderson *et al.*, 2014).

2 Compréhension de la parole avec un apprentissage sans données de référence (Zero-shot)

Le modèle de compréhension employée dans notre étude est celui proposé dans (Ferreira *et al.*, 2015) et présenté dans la figure 1. Il est basé sur une analyse sémantique sans données de référence (Zero-shot Semantic Parser, ZSSP) et fait usage de trois composants principaux. Le premier est un espace sémantique F basé sur une représentation vectorielle compacte de mots apprise avec des

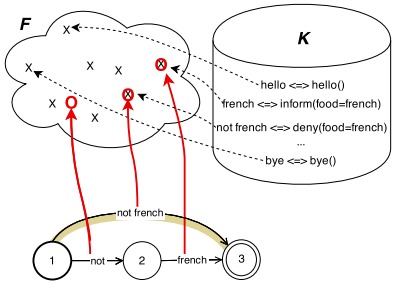


FIGURE 1: Modèle de d'analyseur sémantique sans données de référence (modèle ZSSP)

réseaux de neurones profonds (Mikolov *et al.*, 2013a) à partir de données générales.

Le deuxième composant est une base de connaissances K qui peut être vue comme un dictionnaire d'exemples dans F utilisé pour relier l'espace sémantique à l'espace de sortie du système. Dans K, des coefficients d'affectation mesurent la correspondance entre un exemple et un acte de dialogue connus. Ces valeurs, indiquent le degré de correspondance entre une phrase et une étiquette sémantique. Ces coefficients, initialisés à 1 pour l'ensemble d'exemples/actes extrait de l'ontologie, seront ensuite modifiés en fonction des retours des utilisateurs.

Enfin, un analyseur sémantique qui extrait une liste ordonnée des meilleures hypothèses de séquences d'étiquettes sémantiques à partir d'un transducteur à états finis représentant l'ensemble des hypothèses pour une phrase utilisateur (scorées par des informations issues de F et de K). Par exemple pour la phrase « not french », trois formes de surface différentes sont extraites : « not », « french » et « not french ». Ces formes de surface sont ensuite projetées dans l'espace sémantique F (cercles rouges dans la figure 1) pour être comparées aux vecteurs associés aux exemples de la base de connaissance K (croix noires dans la Fig. 1). Pour ce faire un critère de similarité (e.g. similarité cosinus) entre ces vecteurs est employé. On note que, malgré la présence de K, le ZSSP ne dispose pas de données de référence qui seraient des phrases entières annotées sémantiquement.

3 Optimisation en ligne de la stratégie d'adaptation du modèle

L'objectif principal des travaux présentés dans ce papier est d'enrichir la stratégie d'adaptation en ligne (telle que décrite dans (Ferreira *et al.*, 2015)) pour permettre également la création de concepts/valeurs (extension de domaine) tout en tentant d'optimiser un ratio coût/amélioration du modèle. Dans cette étude, nous adoptons une stratégie simple basée sur un algorithme de bandit contre un adversaire pour résoudre le problème d'optimisation de la stratégie d'adaptation du modèle. Avant d'aller plus loin dans la formulation, nous proposons en premier lieu d'expliciter la problématique sur un cas statique (i.e. sans adaptation).

3.1 Cas statique

Nous postulons que le système a le choix entre plusieurs actions vis à vis de l'utilisateur afin d'améliorer la détection automatique des étiquettes sémantiques associées aux énoncés de celui-ci.

Nous devons d'abord définir l'espace d'actions considéré (qui correspondent donc aux bras des bandits dans la littérature relative aux algorithmes de bandits). Toutefois, nous pouvons déjà prévoir que chaque action implique une collaboration plus ou moins importante de la part de l'utilisateur. Par conséquent, nous introduisons une mesure de l'effort utilisateur relatif à l'action effectivement choisie par le système sous la forme d'une fonction de coût. Nous définissons également une mesure de l'inefficacité du modèle que nous tenterons de réduire dans le temps. Cette mesure nous permet de quantifier l'amélioration de modèle résultant d'une action spécifique. Enfin le problème de l'adaptation du modèle est formulé comme un problème d'optimisation linéaire où le système a la pleine connaissance de la fonction objectif.

3.2 Espace d'actions et fonction de coût traduisant l'effort utilisateur

Lorsque l'utilisateur fournit une phrase, le système peut choisir une action (à partir d'une distribution de probabilité) parmi un ensemble $\mathcal{I}$ de M actions. Dans cette configuration préliminaire, nous considérons le cas où $M = 3$ et où $\mathcal{I}$ peut être défini comme :

$$\mathcal{I} := \{\text{Skip, YesNoQuestions, AskAnnotation}\}. \quad (1)$$

Soit $i \in \mathcal{I}$ l'indice de l'action. Nous supposons que l'effort de l'utilisateur (coût de l'action), $\phi(i) \in \mathbb{N}$, peut être mesuré par le nombre d'échanges réalisés entre le système et l'utilisateur pour mener à bien l'action i.

Soit $d \in [0, 1]$ le poids moyen des arcs constituant le meilleur chemin de la machine à états finis utilisée par l'analyseur sémantique . Selon l'action de raffinement choisie i, d est mis à jour en $d'(i) \in [0, 1]$ en raison de la modification du modèle qui en résulte. Une des composantes étant la fonction de coût nous cherchons à résoudre un problème de minimisation globale, l'inefficacité est prise en compte, au lieu de l'efficacité.

Une description des différentes actions et de leurs coûts et inefficacités associés est donnée ci-dessous :
•**Skip :** n'appliquer aucune mise à jour au modèle. Le coût de cette action est toujours considéré comme étant nul ($\phi(\text{Skip}) = 0$) puisque l'utilisateur n'est pas sollicité et la mesure de l'inefficacité reste donc constante, $d'(Skip) = d$.
•**YesNoQuestions :** mettre à jour K avec les réponses oui/non données par l'utilisateur aux questions de confirmation sur les étiquettes sémantiques détectées dans la meilleure hypothèse sémantique. Si cette action est prise, le système tentera en premier lieu une confirmation générale sur la meilleure hypothèse pour un coût de 1. En cas de négation, une confirmation ($+1$ sur le coût) sera demandée pour chaque étiquette sémantique détectée.
Ces évaluations utilisateurs sont converties en un ensemble U de m tuples $U := (c_l, T_l, f_l)_{1 \leq l \leq m}$, où ($c_l$, T_l) est un couple forme-de-surface / étiquette-sémantique proposé à l'utilisateur et f_l est son retour (1 positif, 0 négatif). Compte tenu de K et U après chaque interaction, l'algorithme 1 est utilisé pour mettre à jour K en K^* Ainsi, $d'(YesNoQuestions) = \delta$ où δ est le nouveau poids moyen de l'énoncé récemment actualisé dans K^*.
•**AskAnnotation :** demander à l'utilisateur d'annoter son tour de parole complètement pour mettre à jour K avec de nouveaux exemples positifs. Si cette action est prise, le système tentera en premier lieu une confirmation générale sur la meilleure hypothèse pour un coût de 1. Dans le cas d'un rejet, l'utilisateur devra procéder à l'annotation étiquette par étiquette de l'énoncé. Pour ce faire, nous supposons que pour chacune d'entre elles l'utilisateur informera le système de la localisation dans sa phrase de l'acte de dialogue qu'il s'apprête à annoter ($+1$ sur le coût) puis identifiera consécutivement

l'$acttype$, le $concept$ et la $valeur$.
Par cette action de nouveaux concepts et valeurs pourront donc être ajoutés par l'utilisateur au module SLU (extension du domaine). Si l'hypothèse sémantique de l'énoncé est validée dans son intégrité par l'utilisateur nous considérons que tous les couples forme-de-surface / étiquette-sémantique extraits de la meilleure hypothèse sémantique (plus court chemin) ont été évalués de façon positive par l'utilisateur. Sinon, les m' couples forme-de-surface/étiquette-sémantique annotés par l'utilisateur sont considérés comme un ensemble de tuples $U := ((c_l, T_l, 1)_{1 \leq l \leq m'})$. Dans ce cas, de nouveaux concepts et valeurs peuvent être ajoutés dans les sorties possibles du modèle (ajout d'une colonne dans K^*). En raison du fait que des parties de l'énoncé sont désormais dans K^* en tant qu'exemples positifs, $d'(AskAnnotation) \approx 0$.

Finalement, nous devons définir une fonction de perte telle que le système, par le fait de chercher à l'optimiser, réduira dans le même temps la mesure de l'inefficacité actualisée $d'(i)$ et la mesure de l'effort de l'utilisateur $\phi(i)$. Ainsi, nous proposons de définir la fonction de perte $l(i) \in [0, 1]$ comme étant la combinaison convexe des deux mesures précédemment introduites :

$$l(i) := \underbrace{\gamma d'(i)}_{\substack{\textbf{amélioration} \\ \textbf{du modèle}}} + \underbrace{(1-\gamma)\frac{\phi(i)}{\phi_{max}}}_{\textbf{effort utilisateur}} \tag{2}$$

où $\gamma \in [0, 1]$ permet de régler l'importance de l'amélioration du modèle sur l'effort utilisateur dans le processus d'optimisation. $\phi_{max} \in \mathbb{N}_+$ correspond au nombre maximal d' échanges possibles entre le système et l'utilisateur (dans un même tour de dialogue).

Soit $\mathbf{p} \in \Delta(3) := \{\mathbf{q} \in \mathbb{R}_+^3 \mid \sum_{i \in \mathcal{I}} q(i) = 1\}$ la distribution de probabilité sur les différentes actions. L'objectif d'adaptation du modèle est donc défini comme :

$$\min_{\mathbf{p} \in \Delta(3)} E[l] = \sum_i p(i) l(i). \tag{3}$$

Si nous avions une pleine connaissance de $l(i)$ pour chaque action i, le problème d'adaptation du modèle serait équivalent à celui consistant à résoudre $\min_i\{l(i)\}$. Cependant, dans le scénario considéré, ce cadre ne peut pas être appliqué car la fonction de perte $l(i)$ n'est pas connue explicitement (pas observable pour toutes les actions à tous les instants). De ce fait, les algorithmes de bandit sont les plus adaptés.

3.3 Cas du bandit contre un adversaire

Pour le problème d'adaptation du modèle par une méthode de bandit contre un adversaire les paramètres connus sont l'espace d'actions $\mathcal{I}$ et le coefficient $\gamma \in [0, 1]$. À chaque tour $t = 1, 2, \ldots$, le système reçoit un énoncé utilisateur, en extrait la meilleure hypothèse sémantique et obtient d_t puis choisit une action $i_t \in \mathcal{I}$, éventuellement en ayant recours à une action aléatoire (exploration).

Une fois l'action i_t exécutée, le système calcule la nouvelle mesure d'inefficacité $d'_t(i_t)$; l'effort utilisateur à t, $\phi_t(i_t)$, qui correspond au nombre d'échanges effectivement réalisés entre le système et l'utilisateur lors de la réalisation de i_t et la fonction de perte :

$$l_t(i_t) = \gamma d'_t(i_t) + (1-\gamma)\phi_t(i_t)$$

Algorithm 1 Bandit contre un Adversaire, l'algorithme Exp3

Sachant : $\gamma' \in [0, 1]$
Initialiser les poids $w_i(1) = 1$ pour $i = 1, \ldots, M$.
for chaque tour t **do** :
    - calculer $p_i(t) = (1 - \gamma')\frac{w_i(t)}{\sum_{j=1}^{M} w_j(t)} + \frac{\gamma'}{M}$ pour chaque i.
    - déterminer la prochaine action i_t aléatoirement selon la distribution $p_i(t)$.
    - observer la récompense $x_{i_t}(t)$.
    - calculer la récompense estimée $\hat{x}_{i_t}(t) = x_{i_t}(t)/p_{i_t}(t)$.
    - mettre à jour les poids :
    $w_{i_t}(t + 1) = w_{i_t}(t)e^{\gamma' \hat{x}_{i_t}(t)/M}$ et $w_j(t + 1) = w_j(t)$ pour tout autre action j.

Le but sera de trouver $i_1, i_2, \ldots$ tels que pour chaque T, le système minimise la perte cumulée :

$$\sum_{t=1}^{T} l_t(i_t) = \gamma \sum_{t=1}^{T} d'_t(i_t) + (1 - \gamma) \sum_{t=1}^{T} \phi_t(i_t).$$

Aucune hypothèse n'est formulée sur $d'_t(i_t) \in [0, 1]$ et $\phi_t(i_t) \in [0, 1]$. Ainsi, nous ne présupposons pas de l'effet qu'a une action i_{t-l}, avec $l \in \{1, \ldots, t - 1\}$, sur la fonction de perte pour le tour t. Ce choix est justifié par le fait qu'une phrase utilisateur ne peut pas être prédite avec précision par le système sans connaissances a priori robustes.

Parmi les algorithmes de la littérature, nous avons retenu l'algorithme Exp3 (Auer *et al.*, 2002) (voir algorithme 1). Il s'agit d'un algorithme efficace lorsqu'un petit nombre de bras est en jeu. Une preuve mathématique des performances relativement élevées de cet algorithme est notamment donnée dans (Bubeck & Cesa-Bianchi, 2012).

3.4 Simulation

Afin de tester l'algorithme d'apprentissage de la politique d'adaptation du modèle, nous avons choisi dans ces travaux de simuler les réponses de l'utilisateur. Pour ce faire, nous avons mis en place un indicateur à même de déterminer la qualité de la meilleure proposition du ZSSP en fonction d'une référence. En raison du fait que les étiquettes sémantiques $acttype(concept = valeur)$ n'étaient pas alignées aux mots dans le corpus considéré (ici DSTC2) et sachant que ce dernier est une condition préalable pour pouvoir simuler l'annotation en séquence de couples forme-de-surface/étiquette-sémantique nous avons dû donc au préalable procéder à un alignement automatique similaire à celui proposé dans (Huet & Lefèvre, 2011). Ainsi, à chaque tour, nous avons suffisamment d'informations pour être en mesure de répondre avec précision à l'action de la machine (séquences d'actes de dialogue de référence et leurs alignements aux mots). Ici, un sous-ensemble de transcriptions de l'ensemble d'apprentissage de DSTC2 (750 transcriptions d'énoncés utilisateur) est exploité pour évaluer le modèle d'adaptation en ligne.

Dans notre configuration expérimentale, un utilisateur simulé est employé pour répondre aux actions d'adaptation du modèle pour chaque tour de parole dans le dialogue d'origine. Cet utilisateur peut faire usage de trois actions distinctes : *Affirm*, *Negate* et *Inform*. Les actions *Affirm* et *Negate* sont employées pour répondre aux demandes de confirmation liées à l'application des actions d'adaptation

du modèle (AskAnnotation et YesNoQuestions). L'action *Inform* est utilisée exclusivement dans les échanges supplémentaires ayant lieu dans le cadre de l'action système AskAnnotation (par exemple *Inform(acttype=request)*, *Inform(boundaries="austrian food")*). Ici, nous supposons que les sous-dialogues d'annotation peuvent être gérés par un système réel avec un niveau de précision élevé (par exemple en utilisant une grammaire bien calibrée et une logique d'interaction finement réglée). Bien sûr cette hypothèse devra être confirmée en pratique.

4 Expériences et résultats

Notre étude expérimentale a été menée sur une tâche de compréhension de la parole en utilisant les données de la seconde campagne d'évaluation Dialog State Tracking Challenge (DSTC2) (Henderson *et al.*, 2014) qui couvre le domaine de la recherche d'information à propos de restaurants. Nous exploitons ces données (transcriptions, annotations sémantiques...) comme corpus d'apprentissage et de test pour évaluer notre approche d'apprentissage en ligne pour l'étiquetage sémantique. Ainsi, les transcriptions du corpus de test (9890 énoncés utilisateur) seront utilisées pour le test. Un sous-ensemble des transcriptions du corpus d'apprentissage (1472 énoncés) seront exploitées dans notre modèle de raffinement en ligne du modèle.

La configuration du modèle sans données de référence utilisé pour appliquer notre méthode d'apprentissage en ligne correspond à celle présentée dans (Ferreira *et al.*, 2015). Un modèle *word2vec* (Mikolov *et al.*, 2013a) a été utilisé pour apprendre une représentation vectorielle des mots avec 300 dimensions. Ce modèle a été appris avec l'algorithme *Skip-gram* (avec une fenêtre de 10 mots et un softmax hiérarchique) sur une grande quantité de données disponibles librement et présentant une grande couverture thématique.

Ce type de représentation présente certaines régularités avec les propriétés syntaxiques et sémantiques des mots comme celles montrées dans (Mikolov *et al.*, 2013b) ainsi qu'une structure linéaire permettant la combinaison des représentations des mots par une simple addition vectorielle élément par élément. Cette technique est donc utilisée pour projeter nos formes de surface vers leur représentation sémantique vectorielle de type word2vec vue comme une somme des représentations individuelles de chaque mot les constituant.

La base de connaissance utilisée pour notre expérience est initialisée grâce aux descriptions ontologiques fournies dans le cadre du DSTC2 (e.g. listes des concepts/valeurs) ainsi que d'un ensemble d'informations génériques. La sémantique du domaine est constituée de 8 concepts et 215 valeurs. Au total 16 *acttype* sont considérés, il en résulte donc 663 étiquettes sémantiques possibles. Nous avons définis manuellement 53 formes de surface associées aux différents *acttype*s. Par exemple « say again » est utilisé pour représenter l'acte *repeat()*.

Du fait que la technique Exp3 emploie une certaine forme d'exploration stochastique (ici $\gamma' = 0,2$) nous utiliserons une moyenne faite à partir de 20 processus indépendants d'apprentissage en ligne.

La figure 2 donne l'évolution de la probabilité $p_i(t)$ associée à chaque action i telle qu'estimée par l'algorithme Exp3 (avec $\gamma = 0,5$). Nous pouvons observer que chaque action est sélectionnée avec une probabilité comparable au début de la procédure d'optimisation, Exp3 explore. Puis, à mesure que le nombre de tours considérés augmente, on observe que l'influence des deux actions YesNoQuestions et Skip croît. On remarque cependant un avantage clair à l'action Skip lorsqu'il devient plus difficile d'obtenir de nouvelles informations eu égard au coût impliqué pour les collecter.

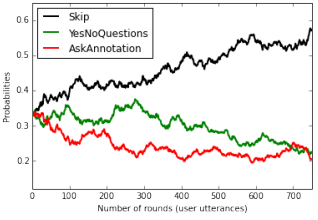


FIGURE 2: Distribution de probabilité estimée par Exp3 au cours du temps sur les différentes actions.

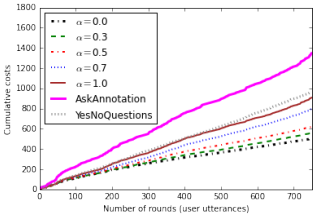


FIGURE 3: Impact de γ sur l'effort utilisateur (coûts) cumulé.

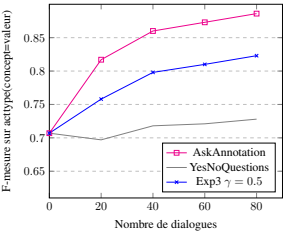


FIGURE 4: Impact du nombre de dialogues employés sur les différentes techniques d'adaptation en ligne en terme de F-mesure.

Dans la figure 3 on compare l'effet de γ sur la stratégie apprise par Exp3 en terme d'effort utilisateur cumulé. Les stratégies **AskAnnotation** et **YesNoQuestions** (stratégies réalisant la même action à chaque tour) sont introduites ici à des fin de comparaison comme méthodes de référence. Nous considérons les performances pour $\gamma \in \{0,0,3,0,5,0,7,1\}$. Nous pouvons observer que la stratégie **AskAnnotation** est la plus coûteuse, suivie par **YesNoQuestions**. Faire varier le paramètre γ semble donc avoir l'effet escompté sur l'apprentissage de la stratégie d'adaptation. Ainsi, plus γ est grand, moins le coût a un impact sur l'apprentissage. De ce fait, lorsque celui-ci est totalement ignoré dans la fonction de perte ($\gamma = 1,0$), l'algorithme Exp3 a tendance à favoriser les actions les plus coûteuses car elles permettent de réduire significativement la mesure d'inefficacité du modèle. Ainsi, γ offre un moyen simple et direct pour régler le compromis entre l'effort de l'utilisateur et l'efficacité du modèle pour une application donnée.

Enfin, dans la figure 4 Exp3 ($\gamma = 0,5$) est comparé à **Askannotation** et **YesNoQuestion** en terme de F-mesure sur les transcriptions du corpus de test. Comme prévu **AskAnnotation** obtient les meilleures performances. En effet, l'utilisation des nouvelles annotations permet au modèle ZSSP de couvrir dynamiquement des actes de dialogue supplémentaires grâce à la mise à jour de K avec des exemples robustes. Du fait que l'objectif de l'algorithme Exp3 est de trouver un compromis entre réduire l'effort de l'utilisateur et l'efficacité du modèle, cette méthode est capable d'atteindre à plus faible coût des performances proches de celles obtenues avec **AskAnnotation** et bien meilleures que celles observées pour **YesNoQuestion** (cette dernière ne pouvant pas capturer de nouveaux concepts).

5 Conclusion

Dans ce papier une approche de bandit contre un adversaire a été employée pour optimiser la stratégie d'adaptation d'un modèle d'analyse sémantique sans données de références et permettre de résoudre le problème d'une couverture initiale limitée sur la sémantique de domaine spécifique. Il a été montré que cette technique est efficace et à même de fournir un moyen pratique de formaliser un compromis entre l'effort de supervision fourni par l'utilisateur et l'amélioration de l'efficacité du système. La généralisation de l'approche d'optimisation proposée ainsi qu'une comparaison plus poussée avec d'autres algorithmes de bandit fera l'objet de futurs travaux.

Références

AUER P., CESA-BIANCHI N., FREUND Y. & SCHAPIRE R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, p. 48–77.

BAYER A. & RICCARDI G. (2013). On-line adaptation of semantic models for spoken language understanding. In *ASRU*.

BUBECK S. & CESA-BIANCHI N. (2012). Regret analysis of stochastic and nonstochastic multiarmed bandit problems. *Foundations and Trends in Machine Learning*, **5**(1), 1–122.

CELIKYILMAZ A., TUR G. & HAKKANI-TUR D. (2011). Leveraging web query logs to learn user intent via bayesian latent variable model. In *ICML*.

DEORAS A. & SARIKAYA R. (2013). Deep belief network based semantic taggers for spoken language understanding. In *INTERSPEECH*.

FERREIRA E., JABAIAN B. & LEFÈVRE F. (2015). Online adaptative zero-shot learning spoken language understanding using word-embedding. In *ICASSP*.

GOTAB P., DAMNATI G., BÉCHET F. & DELPHIN-POULAT L. (2010). Online slu model adaptation with a partial oracle. In *INTERSPEECH*.

HAHN S., DINARELLI M., RAYMOND C., LEFÈVRE F., LEHNEN P., DE MORI R., MOSCHITTI A., NEY H. & RICCARDI G. (2010). Comparing stochastic approaches to spoken language understanding in multiple languages. *IEEE TASLP*, **19**(6), 1569–1583.

HECK L. & HAKKANI-TUR D. (2012). Exploiting the semantic web for unsupervised spoken language understanding. In *SLT*.

HENDERSON M., THOMSON B. & WILLIAMS J. (2014). The second dialog state tracking challenge. In *SIGDIAL*.

HUET S. & LEFÈVRE F. (2011). Unsupervised alignment for segmental-based language understanding. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*.

JABAIAN B., BESACIER L. & LEFÈVRE F. (2013). Comparison and Combination of Lightly Supervised Approaches for Language Portability of a Spoken Language Understanding System. *IEEE TASLP*, **21**(3), 636–648.

LEFÈVRE F., MOSTEFA D., BESACIER L., ESTEVE Y., QUIGNARD M., CAMELIN N., FAVRE B., JABAIAN B. & ROJAS-BARAHONA L. (2012). Robustness and portability of spoken language understanding systems among languages and domains : the PORT-MEDIA project. In *LREC*.

LORENZO A., ROJAS-BARAHONA L. & CERISARA C. (2013). Unsupervised structured semantic inference for spoken dialog reservation tasks. In *SIGDIAL*.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

MIKOLOV T., YIH W. & ZWEIG G. (2013b). Linguistic regularities in continuous space word representations. In *NAACL-HLT*.

RALAIVOLA L., FAVRE B., GOTAB P., BÉCHET F. & DAMNATI G. (2011). Applying multiclass bandit algorithms to call-type classification. In *Workshop on Automatic Speech Recognition & Understanding, ASRU*.

TUR G., HAKKANI-TUR D., HILLARD D. & CELIKYILMAZ A. (2011). Towards unsupervised spoken language understanding : Exploiting query click logs for slot filling. In *INTERSPEECH*.

Patrons Rythmiques et Genres Littéraires en Synthèse de Parole

Elisabeth Delais-Roussarie[1], Damien Lolive[2], Hiyon Yoo[1] et David Guennec[2]
(1) LLF - UMR 7110 & Université Paris-Diderot, France
(2) IRISA - UMR 6074, Université Rennes 1, France
elisabeth.roussarie@wanadoo.fr, damien.lolive@irisa.fr,
yoo@linguist.univ-paris-diderot.fr, david.guennec@irisa.fr

RÉSUMÉ

Ces vingt dernières années, la qualité de la parole synthétique s'est améliorée grâce notamment à l'émergence de nouvelles techniques comme la synthèse par corpus. Mais les patrons rythmiques obtenus ne sont pas toujours perçus comme très naturels. Dans ce papier, nous comparons les patrons rythmiques observés en parole naturelle et synthétique pour trois genres littéraires. Le but de ce travail est d'étudier comment le rythme pourrait être amélioré en synthèse de parole. La comparaison des patrons rythmiques est réalisée grâce à une analyse de la durée relativement à la structure prosodique, les données audio provenant de six comptines, quatre poèmes et deux extraits de conte. Les résultats obtenus laissent penser que les différences rythmiques entre parole naturelle et synthétique sont principalement dues au marquage de la structure prosodique, particulièrement au niveau des groupes intonatifs. De fait, le taux d'allongement des syllabes accentuées en fin de groupes intonatifs est beaucoup plus important en synthèse que dans la parole naturelle.

ABSTRACT

How to improve rhythmic patterns according to literary genre in synthesized speech *.

In the last twenty years, the quality of synthesized speech has greatly improved with the emergence of new TTS techniques, including corpus-based synthesis systems. Yet the rhythmic patterns obtained do not always sound very natural. In this paper, we compare the rhythmic patterns observed in natural and synthesized speech for three literary forms. The aim of the study is to evaluate how rhythm could be improved in synthesized speech. The comparison of the rhythmic patterns is done by analyzing duration in relation to prosodic structure on a set of texts (six rhymes, four poems and two extracts from fairy tales). This approach allows showing that rhythmic differences between synthesized and natural speech are mostly due to the marking of prosodic structure, especially at the level of the intonational phrase. The lengthening rate for accented syllables located at the end of IPs is much more important in synthesized speech than in natural speech.

MOTS-CLÉS : Patrons rythmiques, phonogenre, synthèse de la parole, structure prosodique.

KEYWORDS: Rhythmic patterns, phono-genre, speech synthesis, prosodic structure..

1 Introduction

Ces dernières décennies, la qualité globale de la parole synthétisée s'est améliorée de façon notable avec l'émergence de nouvelles techniques de synthèse comme la synthèse par corpus (Sagisaka, 1988; Hunt & Black, 1996). Néanmoins, générer une prosodie naturelle qui tienne compte des genres et

∗. Cet article est tiré d'une publication par les mêmes auteurs à la conférence Speech Prosody 2016.

styles de parole reste un challenge (Schröder, 2009; Obin, 2011), en particulier pour les aspects rythmiques. De fait, la composante rythmique semble souvent peu naturelle en synthèse de parole et doit être améliorée pour permettre une meilleure utilisation de la synthèse dans de nombreuses applications (jeu vidéo, logiciel éducatif, lecture de livres audio, etc.).

Dans un projet de recherche visant à utiliser la synthèse de parole pour favoriser l'apprentissage de l'écriture à des enfants de cycle 2 (CP, CE1), il fallait améliorer le système de synthèse de parole afin qu'il puisse lire de façon claire et naturelle des contes, des poèmes et des comptines. Pour tenter d'atteindre cet objectif, nous avons comparé les patrons rythmiques obtenus en parole naturelle et en parole synthétique pour chacun des genres littéraires visés (comptines, poèmes, contes). Nous avions émis au départ l'hypothèse que les patrons rythmiques les plus précis seraient observés pour les contes, les corpus utilisés pour extraire les unités de parole lors de la synthèse contenant essentiellement des textes lus comparables à des récits. Cependant, les résultats obtenus n'ont pas confirmé cette hypothèse, les lectures de comptines étant souvent plus satisfaisantes. Aussi, avons-nous tenté de comprendre pourquoi les patrons rythmiques sont plus adéquats dans le cas des poèmes et des comptines que dans le cas des contes, alors que les corpus utilisés pour générer les stimuli ne contenaient pas ce genre littéraire.

Cet article est organisé de la manière suivante. La section 2 fournit une description des données et de la méthode utilisée. Dans la section 3, les résultats obtenus grâce à la comparaison des patrons de durée et de débit de parole dans les deux types de parole (parole naturelle vs. parole synthétisée) sont présentés pour chaque genre. Ces résultats sont ensuite discutés dans la section 4, le but de cette discussion étant essentiellement de voir comment améliorer les systèmes de synthèse.

2 Corpus et méthodologie

2.1 Corpus

Le corpus utilisé pour étudier les patrons rythmiques en parole naturelle et synthétique est constitué de trois types distincts de textes adressés à des enfants : six comptines, quatre poèmes et deux extraits de contes. Le tableau 1 présente la composition quantitative du corpus par genre littéraire. Les différences de réalisation entre les locuteurs (qui sont indirectement notées par l'écart entre le nombre de syllabes par genre à multiplier par cinq, c'est-à-dire le nombre de locuteurs, et le nombre effectif de syllabes) résultent principalement de l'insertion ou de l'élision de schwas, ou de l'omission d'un mot. Le nombre effectif de syllabes obtenu pour la parole synthétique est donné entre parenthèses.

Genre littéraire	Nombre de mots	Nombre de syll.	Nombre de syll. effectif
Comptines	158	228	1137 (454 for synth.)
Poèmes	290	422	2155 (808 for synth.)
Contes	522	777	3861 (1538 for synth.)
Total	970	1427	7153 (2800)

TABLE 1 – Composition du corpus

L'ensemble des textes a été produit par cinq voix différentes (deux voix synthétiques et trois voix naturelles). Pour les voix naturelles, le corpus a été enregistré par trois locuteurs (deux hommes et une femme) dans un studio d'enregistrement. Les participants ont eu le temps de lire les textes et de

les répéter avant l'enregistrement. Parmi ces locuteurs, deux ont lu les textes de la même manière que des parents liraient une histoire à leurs enfants, tandis que le troisième est un acteur confirmé et les a lus avec beaucoup plus d'expressivité.

Les stimuli synthétisés ont été produits grâce au système de synthèse par corpus présenté dans (Guennec & Lolive, 2014), pour lequel des filtres de pré-sélection sont utilisés en lieu et place d'un coût cible. Pour cette étude, les filtres ordonnés utilisés au niveau du phonème sont les suivants :

1. Identité dans les étiquettes associées aux segments (obligatoire).
2. Nature de l'unité : phonème ou autre ? (obligatoire)
3. Est-ce que le segment/phonème est dans la dernière syllabe de la phrase ?
4. Le segment se situe-t-il dans la dernière syllabe d'un groupe prosodique majeur (IP) ?
5. Le segment est-il dans la dernière syllabe d'un mot ?
6. Le segment est-il dans une syllabe réalisée avec une intonation montante ?

Lors de la recherche, si le nombre d'unités correspondant à un ensemble de filtres est insuffisant, le dernier filtre de l'ensemble est relâché. Cela permet d'élargir le champ de recherche en réduisant le nombre de contraintes appliquées. Dans tous les cas, les deux premiers filtres sont toujours utilisés. De manière complémentaire, une pénalité est appliquée pour les classes de phonèmes pour lesquelles la concaténation semble risquée (Alain *et al.*, 2015). Par exemple, une concaténation sur une voyelle est plus sujette à l'apparition d'un artéfact audible qu'une concaténation réalisée sur la partie silencieuse d'une plosive ou même sur une fricative.

Concernant la prosodie, aucun traitement spécifique n'est réalisé, et les seules contraintes susceptibles d'améliorer le rythme de la parole générée sont les filtres de pré-sélection. De fait, certains d'entre eux permettent d'appliquer aux unités sélectionnées des contraintes positionnelles comme par exemple la fin d'énoncés ou la fin de mot. En ce qui concerne les pauses, elles sont placées de manière arbitraire, et une pause de durée fixe est insérée après chaque marque de ponctuation.

Comme nous l'avons déjà mentionné, deux voix de synthèse ont été utilisées pour cette étude :

— la voix d'homme SY-P, construite à partir de 10 heures de parole extraites d'un livre audio (un roman lu par un acteur) ;
— la voix de femme SY-A, construite à partir de 7 heures de parole lue, les éléments lus ayant été sélectionnés spécifiquement pour la construction d'un système de synthèse. Les différences de contenu et de taille des corpus amènent à considérer SY-P comme une voix plus expressive que SY-A, qui est plus neutre.

Pour générer les stimuli de synthèse, la structure des strophes et des vers dans les poèmes et comptines a été représentée par des marques de ponctuation. Ainsi, pour obtenir la version synthétisée, les trois strophes sous (1), extraites du poème "La fourmi" de R. Desnos ont été mises en forme comme indiqué sous (2).

(1) *Une fourmi traînant un char*
plein de pingouins et de canards
ça n'existe pas, ça n'existe pas

Une fourmi parlant français
parlant latin et javanais
ça n'existe pas, ça n'existe pas

eh ! et pourquoi pas !

(2) Une fourmi traînant un char, plein de pingouins et de canards, ça n'existe pas, ça n'existe pas. Une fourmi parlant français, parlant latin et javanais, ça n'existe pas, ça n'existe pas. Eh ! Et pourquoi pas !

Comme on peut le voir, la fin des strophes est systématiquement indiquée par un point même s'il n'y avait pas de ponctuation dans le texte original. La fin des vers est retranscrite par une virgule, sauf dans le cas où une ponctuation existait déjà.

2.2 Méthodologie

Les enregistrements audio ont d'abord été transcrits et segmentés en phrases sous PRAAT (Boersma & Weenink, 2001). La transcription orthographique a ensuite été phonétisée, et le signal acoustique automatiquement segmenté en phones, syllabes, et mots avec EASYALIGN (Goldman, 2011). Les transcriptions phonétiques et les segmentations acoustiques ont été vérifiées et corrigées, si nécessaire. L'ensemble des données a été utilisé pour l'analyse rythmique et prosodique.

La voyelle plutôt que la syllabe a été choisie comme unité de base pour générer les patrons de durée et pour analyser et comparer les durées des pauses et les débits en fonction des genres et des locuteurs. Ce choix résulte du fait que les structures syllabiques varient beaucoup en français (entre, par exemple, des syllabes de forme CCVC et d'autres de forme CV). La durée des syllabes ne constituent donc pas un indicateur robuste pour évaluer les taux d'allongement. Comme le nombre de voyelles par contextes prosodiques était limité en raison de la taille du corpus, aucune normalisation des durées n'était possible. Aussi avons-nous décidé de faire une distinction entre voyelles courtes et voyelles longues, même si cette distinction n'existe pas dans le système phonologique du français. Les voyelles nasales ([ɔ̃],[ɑ̃], [ɛ̃] et [œ̃]) et les séquences composées d'une semi-voyelle et d'une voyelle en position de noyau (par exemple, [jɛ̃] dans *tiens* [tjɛ̃], [wa] dans *noir* [nwaʁ]) ont été codées comme des voyelles longues, tandis que les autres voyelles ont été considérées comme courtes.

De nombreux travaux consacrés à la prosodie du français ont montré que l'intonation et l'accentuation sont très liées dans cette langue (cf. (Post, 2011)) ; aussi avons-nous décidé de partir des découpages prosodiques pour étudier les schémas rythmiques. Les différents textes ont donc été segmentés en groupes prosodiques, une distinction étant faite entre trois niveaux de structuration : le mot prosodique (MP) qui correspond à un mot lexical précédé des mots grammaticaux qui en dépendent, le syntagme phonologique (SP) qui est borné à droite par une tête lexicale de projection syntagmatique maximale et le groupe intonatif (IP). Pour pouvoir comparer les données malgré les différences possibles de réalisation et pour éviter une certaine circularité, nous avons décidé de dériver les unités prosodiques à partir du texte, et plus précisément des informations morpho-syntaxiques, voir (Delais-Roussarie, 1996; Martin, 1987; Pasdeloup, 1992). De plus, comme la dernière syllabe des groupes prosodiques est considérée en français comme accentuée et est habituellement allongée (Fletcher, 1991), trois catégories de syllabes accentuées ont été retenues pour comparer les taux d'allongement par rapport à la position prosodique :

— AC-MP correspond à la dernière syllabe accentuée d'un mot prosodique, c'est-à-dire d'un mot d'une catégorie lexicale tels que le verbe V, le nom N, l'adjectif A ou l'adverbe Adv (voir, entre autres, (Mertens *et al.*, 2001; Nespor & Vogel, 1986)) ;
— AC-SP coïncide avec la dernière syllabe accentuée d'un syntagme phonologique, c'est-à-dire la dernière syllabe accentuée d'une tête lexicale de projection syntaxique (voir, par exemple, (Delais-Roussarie, 1996; Post, 2000; Selkirk, 1986)) ;

— AC-IP correspond à la dernière syllabe accentuée d'un IP, les frontières d'IP étant localisées à la fin des clauses, des constituants syntaxiques détachés, ou, dans les poèmes et les comptines, des vers (voir, entre autres, (Nespor & Vogel, 1986; Delais-Roussarie *et al.*, 2015; Portes & Bertrand, 2011)).

3 Résultats

L'étude des durées observées pour les voix naturelles et synthétiques a permis de comparer les débits de parole, la durée et la distribution des pauses, et le marquage de la structure prosodique.

3.1 Débit de parole et pauses

La durée totale des textes lus a été utilisée pour calculer, pour chaque locuteur et pour chaque genre, le débit de parole, le taux d'articulation et les durées des pauses. Les différences entre débit de parole et taux d'articulation reposent sur le fait que les pauses ne sont pas prises en compte dans le second cas (Simon *et al.*, 2010). Le tableau 2 présente les résultats obtenus pour chaque locuteur dans les trois genres. Pour chacun d'entre eux, les deux premières lignes concernent le débit de parole et le taux d'articulation, tandis que les deux dernières lignes portent sur la durée.

Comptines	LOD	DRE	GOR	SY-A	SY-P
Débit de parole moyen (ph./sec.)	9.9	7.35	7.08	7.63	9.09
Taux d'articulation moyen (ph./sec)	12.09	7.83	8.53	9.79	12.61
Durée totale des pauses (ms)	2178.92	1449.22	2573.76	3025	3000
% de pause moyen	25.27	13.60	24.15	29.06	33.80
Poèmes	LOD	DRE	GOR	SY-A	SY-P
Débit de parole moyen (ph./sec.)	10.6	8.16	6.28	8.26	9.45
Taux d'articulation moyen (ph./sec)	13.60	9.32	8.72	10.70	12.85
Durée totale des pauses	1534	1373.36	2590.52	2000	2000
% de pause moyen	27.38	18.10	33.85	28.17	31.29
Contes	LOD	DRE	GOR	SY-A	SY-P
Débit de parole moyen (ph./sec.)	10.58	8.74	8.18	9.31	10.79
Taux d'articulation moyen (ph./sec)	14.99	10.08	11.09	11.36	13.68
Durée totale des pauses	1331.33	763.40	1482.06	992	992.14
% de pause moyen	32.82	18.06	29.96	21.96	24.79

TABLE 2 – Débit de parole et taux d'articulation en phones/sec, durée et pourcentage de pauses (relativement à la durée totale de lecture).

Les taux d'articulation et les débits de parole observés pour chaque genre varient de manière importante, mais on ne peut pas dire que les voix de synthèse diffèrent des voix naturelles : LOD et SY-P possèdent pour les trois genres considérés un débit plus rapide que les locuteurs SY-A, GOR et DRE (qui ont des débits plus lents, mais relativement semblables). Si on compare pour un genre donné les débits des différentes voix, on s'aperçoit que les taux d'articulation obtenus par la synthèse sont dans les limites de ceux observés pour les voix naturelles.

Une comparaison inter-genres montre que les locuteurs adaptent leur débit de parole et leur taux d'articulation en fonction du genre, des débits plus lents étant mis en œuvre pour la lecture des comptines et des poèmes. Cette adaptation est, comme on s'y attendait, moins claire pour la parole synthétique. De fait, pour une voix donnée et pour tous les genres, le même corpus et la même procédure de sélection d'unités sont utilisés. Néanmoins, les différences entre voix naturelles et voix synthétiques demeurent mineures, ce qui signifie que l'adaptation découle également de la composition interne des textes.

Concernant les pauses, une différence importante existe entre parole naturelle et parole synthétique dans tous les genres. La proportion de pauses est moins importante dans les comptines que dans les contes pour les trois locuteurs ; en revanche, il y a plus de pauses dans les comptines que dans les contes pour les deux voix de synthèse. Vus les mécanismes de placement des pauses utilisés par le synthétiseur, ces résultats sont tout à fait logiques. De plus, en parole naturelle, la durée des pauses semble dépendre du taux d'articulation (la proportion de pauses est en effet moins importante lorsque le débit est lent, comme par exemple dans les comptines et les poèmes) ; mais une telle corrélation n'apparaît pas en synthèse, une durée fixe étant assignée aux pauses en fonction de la force de la frontière prosodique.

Dans l'ensemble, on n'observe pas de grosses différences entre parole naturelle et synthétique pour le débit de parole et le taux d'articulation. En effet, les voix de synthèse et les voix naturelles varient dans les mêmes proportions. En revanche, pour la durée et la proportion des pauses, il existe des différences entre la synthèse et les voix naturelles.

3.2 Structure prosodique et durée

En règle générale, les allongements indiquent en français le phrasé et l'accentuation. Les syllabes accentuées, qui correspondent à la dernière syllabe pleine à chaque niveau de structuration prosodique, sont allongées, leur taux d'allongement étant proportionnel à la force de la frontière prosodique, voir, entre autres, (Post, 2000; Portes & Bertrand, 2011). Aussi avons-nous voulu vérifier si cela se retrouve dans les voix de synthèse. Pour ce faire, les taux d'allongement ont été calculés en comparant les durées des voyelles dans les syllabes non accentuées aux durées des segments vocaliques dans les syllabes accentuées, et cela à tous les niveaux de hiérarchie prosodique (mot prosodique, syntagme phonologique et groupe intonatif). Le tableau 3 présente les résultats obtenus par genre.

Il existe une variation relativement importante de la durée des voyelles non accentuées dans les différents genres, et cela pour les trois voix naturelles. De manière générale, les voyelles en position non accentuée sont plus longues dans les comptines et les poèmes que dans les contes. Par comparaison, aucune variation claire n'est observée entre genres pour les voix synthétisées. Ce résultat confirme le fait que les locuteurs adaptent leur débit de parole en fonction du genre, ce que ne fait pas la synthèse de parole.

En ce qui concerne le marquage de la structuration prosodique, des allongements se produisent toujours à la fin des groupes prosodiques à tous les niveaux (mot prosodique, syntagme phonologique et groupe intonatif), en synthèse comme en parole naturelle. Pour tous les genres et tous les locuteurs, les taux d'allongement varient

— de 10 à 40%, au niveau du mot prosodique, avec une moyenne aux alentours de 20% ,
— de 20 à 100% au niveau du syntagme phonologique, avec une moyenne à 40%,
— de 60 à 190% au niveau du groupe intonatif, avec une moyenne de 98% (avec 77% en parole naturelle et 128% en synthèse).

Les taux moyens (à l'exception des IP en parole synthétique) correspondent à ceux souvent donnés dans les travaux sur les patrons de durée en français (Delais-Roussarie, 1996; Pasdeloup, 1992). Dans les comptines, les taux d'allongement ne permettent pas toujours de clairement distinguer les trois niveaux de structuration, en particulier les mots prosodiques des syntagmes phonologiques. Dans les poèmes, la distinction entre SP et IP n'est pas clairement marquée dans les taux d'allongement chez LOD et GOR. On peut aussi noter que les taux d'allongement qui indiquent les frontières des IP sont plus nettement marqués en synthèse qu'en parole naturelle, dans tous les genres, et plus particulièrement chez SY-A.

Comptines	LOD	DRE	GOR	SY-A	SY-P
Durée moyenne voy. non accentuée	66 ms	130 ms	93 ms	81 ms	68 ms
Taux d'allongement AC-MP	30%	20%	20%	20%	30%
Taux d'allongement AC-SP	20%	20%	50%	30%	10%
Taux d'allongement AC-IP	90%	70%	70%	150%	60%
Poèmes	**LOD**	**DRE**	**GOR**	**SY-A**	**SY-P**
Durée moyenne voy. non accentuée	67 ms	110 ms	95 ms	78 ms	69 ms
Taux d'allongement AC-MP	10%	20%	40%	20%	20%
Taux d'allongement AC-SP	60%	40%	100%	50%	40%
Taux d'allongement AC-IP	60%	70%	80%	190%	80%
Contes	**LOD**	**DRE**	**GOR**	**SY-A**	**SY-P**
Durée moyenne voy. non accentuée	59 ms	99 ms	78 ms	77 ms	65 ms
Taux d'allongement AC-MP	10%	20%	10%	20%	20%
Taux d'allongement AC-SP	20%	40%	50%	40%	30%
Taux d'allongement AC-IP	80%	80%	100%	190%	100%

TABLE 3 – Durées moyennes des voyelles dans les syllabes non accentuées (en ms.) et taux d'allongement (en %) pour les trois niveaux de structuration (MP, SP et IP).

Globalement, les patrons de durée obtenus pour la parole synthétique sont relativement comparables à ceux observés en parole naturelle : les différents niveaux de phrasé sont toujours indiqués par un allongement dont le taux varie, très souvent, proportionnellement à la force de la frontière (Post, 2000; Delais-Roussarie *et al.*, 2015; Delais-Roussarie & Feldhausen, 2014).

4 Discussion

Les comparaisons effectuées ne révèlent pas de différences notables entre parole synthétique et parole naturelle. De fait, les variations qui apparaissent pour le débit de parole et le taux d'articulation ne permettent pas de distinguer la parole naturelle de la parole synthétique. En ce qui concerne les allongements et le marquage des frontières prosodiques, l'analyse montre clairement que des allongements sont réalisés en fin de groupement prosodique dans les deux types de parole (naturelle et synthétique), même si des différences apparaissent dans l'importance des taux d'allongement observés au niveau des IP (plus important en synthèse qu'en parole naturelle). Néanmoins, on peut douter que ces différences expliquent à elles seules le manque de naturel de la parole synthétique. De plus, en écoutant les stimuli synthétisés, nous avons été surpris par la qualité des patrons rythmiques observés dans les comptines, en particulier pour SY-A : ils semblent très naturels en comparaison de ceux obtenus pour les contes. En conséquence, les problèmes de rythme rencontrés en synthèse ne

peuvent pas être attribués à des "sur-allongements" au niveau des IPs.

Les taux d'articulation et le débit d'une part, et le marquage par la durée de la structuration prosodique d'autre part, ne peuvent expliquer le manque de naturel dans les motifs rythmiques. Il faut donc trouver d'autres explications. Deux pistes de recherche méritent selon nous d'être explorées : (i) aucune corrélation entre le débit de parole, la force des frontières et la durée des pauses n'a été observée en parole synthétique, alors que cette corrélation existe en parole naturelle (en français, de nombreuses études ont montré que les groupes prosodiques comme le syntagme phonologique ou le groupe intonatif visent soit à posséder le même nombre de syllabes soit la même durée (Delais-Roussarie, 1996; Martin, 1987; Pasdeloup, 1992; Wioland, 1991), les durées des pauses pouvant alors jouer un rôle important dans la recherche de l'isochronie) ; et (ii) les patrons intonatifs jouent probablement un rôle dans la réalisation des motifs rythmiques : en insérant une virgule à la fin de chaque vers dans les poèmes et les comptines, on a forcé la réalisation d'un contour mélodique non final montant (contour de continuation majeure), ce qui a conduit à la répétition régulière d'une forme mélodique et a renforcé l'impression de rythme, ces résultats laissant penser que la récurrence de motifs mélodiques est cruciale pour le rythme.

5 Conclusion et perspectives

L'analyse des patrons de durée observés en parole naturelle et synthétique dans trois genres littéraires montre clairement que la durée ne peut pas, à elle seule, expliquer le manque de naturel de la synthèse de parole, notamment pour les aspects rythmiques. Les valeurs obtenues pour les durées segmentales et pour le marquage de la structure prosodique sont comparables dans bien des cas. Des travaux complémentaires sur des corpus plus grands sont donc nécessaires. De plus, trois points peuvent être avantageusement intégrés pour améliorer la procédure de sélection d'unités dans le système de synthèse de parole, et par là-même, les patrons rythmiques et prosodiques :

— Distinguer clairement les différents niveaux de structuration prosodique et les prendre en compte dans la sélection des unités, notamment comme contrainte positionnelle ;
— Prendre en compte la forme des mouvements mélodiques réalisés sur les syllabes accentuées : la procédure qui a été utilisée dans les comptines et les poèmes en forçant à l'insertion de contours intonatifs similaires à intervalles réguliers a donné des résultats très satisfaisants ;
— Adapter les taux d'allongements, les taux d'articulation et la durée des pauses en fonction des genres, mais aussi pour rendre compte des corrélations qui existent entre ces éléments.

Remerciements

Le travail présenté ici a été soutenu par l'opération PPC 7 du Labex "Empirical Foundations in Linguistics" (ANR-10-LABX-0083). Il a également bénéficié du soutien financier de l'Agence Nationale de la Recherche dans le cadre du projet ANR SynPaFlex (ANR-15-CE23-0015).

Références

ALAIN P., CHEVELU J., GUENNEC D., LECORVÉ G. & LOLIVE D. (2015). The IRISA Text-To-Speech system for the blizzard challenge 2015. In *Proc. of the Blizzard Challenge 2015 Workshop*.

BOERSMA P. & WEENINK D. (2001). Praat, a system for doing phonetics by computer.

DELAIS-ROUSSARIE E. (1996). *Phonological phrasing and accentuation in French*, In M. NESPOR & N. SMITH, Eds., *Dam Phonology : HIL phonology papers II, den Haag : Holland Academic Graphics*, p. 1–38.

DELAIS-ROUSSARIE E. & FELDHAUSEN I. (2014). "variation in prosodic boundary strength : a study on dislocated XPs in french". In *Proc. of Speech Prosody*, p. 1052—1056.

DELAIS-ROUSSARIE E., POST B., AVANZI M., BUTHKE C., DI CRISTO A., FELDHAUSEN I., JUN S., MARTIN P., MEISENBURG T., RIALLAND A. *et al.* (2015). *Intonational Phonology of French : Developing a ToBI system for French*, In S. FROTA & P. PRIETO, Eds., *Intonation in Romance*, p. 63–100. Oxford University Press.

FLETCHER J. (1991). Rhythm and final lengthening in french. *Journal of phonetics*, **19**(2), 193–212.

GOLDMAN J.-P. (2011). Easyalign : an automatic phonetic alignment tool under praat. In *Proc. of Interspeech*, p. 3233–3236.

GUENNEC D. & LOLIVE D. (2014). Unit selection cost function exploration using an A* based text-to-speech system. In *Proc. of TSD*, p. 432–440 : LNCS, Springer, Heidelberg.

HUNT A. J. & BLACK A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. of ICASSP*, volume 1, p. 373–376.

MARTIN P. (1987). Prosodic and rhythmic structures in french. *Linguistics*, **25**(5), 925–950.

MERTENS P., GOLDMAN J.-P., WEHRLI E. & GAUDINAT A. (2001). La synthèse de l'intonation à partir de structures syntaxiques riches. *Traitement Automatique des Langues*, **42**(1), 145–192.

NESPOR M. & VOGEL I. (1986). Prosodic phonology, vol. 28. *Dordrecht : Foris*.

OBIN N. (2011). *MeLos : Analysis and modelling of speech prosody and speaking style*. PhD thesis, Université Pierre et Marie Curie-Paris VI.

PASDELOUP V. (1992). *A prosodic model for French text-to-speech synthesis : A psycholinguistic approach*, In C. B. G. BAILLY & T. SAWALLIS, Eds., *Talking Machines. Theories, Models and Designs*, p. 335–348. Elsevier Science Publishers.

PORTES C. & BERTRAND R. (2011). Permanence et variation des unités prosodiques dans le discours et l'interaction. *Journal of French Language Studies*, **21**(01), 97–110.

POST B. (2000). Tonal and phrasal structures in french intonation. *The Hague : Holland Academic Graphics*.

POST B. (2011). *The multi-faceted relation between phrasing and intonation contours in French*, In C. GABRIEL & C. LLEÒ, Eds., *Intonational Phrasing in Romance and Germanic : Crosslinguistic and bilingual studies*, p. 44–74. Amsterdam : Benjamins.

SAGISAKA Y. (1988). Speech synthesis by rule using an optimal selection of non-uniform synthesis units. In *Proc. of ICASSP*, volume 1, p. 679–682.

SCHRÖDER M. (2009). Expressive speech synthesis : Past, present, and possible futures. In *Affective information processing*, p. 111–126. Springer.

SELKIRK E. (1986). On derived domains in sentence phonology. *Phonology*, **3**(01), 371–405.

SIMON A.-C., AUCHLIN A., AVANZI M., GOLDMAN J.-P. *et al.* (2010). *Les phonostyles : une description prosodique des styles de parole en français*, In M. ABÉCASSIS & G. LEDEGEN, Eds., *Les voix des français : en parlant, en écrivant*. Bern : Lang.

WIOLAND F. (1991). *Prononcer les mots du français. Des sons et des rythmes*. Paris : Hachette.

Une pénalité floue fondée phonologiquement pour améliorer la sélection d'unité

David Guennec Damien Lolive
IRISA - ENSSAT/Université de Rennes 1, 6 rue de Kerampont, 22305 Lannion Cedex, France
{david.guennec, damien.lolive}@irisa.fr

RÉSUMÉ

Les systèmes de synthèse par corpus reposent, sauf de rares exceptions, sur des coûts cibles et des coûts de concaténation pour sélectionner la meilleure séquence d'unités. Le rôle du coût de concaténation est de s'assurer que l'assemblage de deux segments de parole ne causera l'apparition d'aucun artefact acoustique. Pour cette tâche, des distances acoustiques (MFCC, F0) sont généralement utilisées, mais dans de nombreux cas cela ne suffit pas. Dans cet article, nous introduisons une pénalité héritée du domaine de la couverture de corpus dans le coût de concaténation afin de bloquer certaines concaténations en fonction de la classe phonologique des diphones à concaténer. En outre, une seconde version faisant appel à une fonction floue est proposée pour relâcher la pénalité en fonction du positionnement du coût de concaténation par rapport à sa distribution. Une évaluation objective montre que la pénalité est efficace et amène à un meilleur classement des séquences d'unités candidates au cours de la sélection. Une évaluation subjective révèle une performance supérieure de l'approche floue.

ABSTRACT

A Phonologically Motivated Penalty To Improve Unit Selection

Unit selection speech synthesis systems rely, except for rare cases, on target and concaténation costs for selecting the best unit sequence. The role of the concatenation cost is to insure that joining two voice segments will not cause any acoustic artefact to appear. For this task, acoustic distances (MFCC, F0) are typically used but in many cases, this is not enough. In this paper, we introduce a penalty in the concaténation cost, inherited from the field of corpus covering, in order to block some concaténations based on their phonological class. Moreover, a derived fuzzy version is proposed to relax the penalty based on the concaténation quality with respect to the cost distribution. An objective evaluation showed that the penalty is effective to better rank candidate unit sequences during selection. The subjective evaluation we conducted reveals a superior performance of the fuzzy approach.

MOTS-CLÉS : Coût de Concaténation, Synthèse Par Corpus, Sélection d'Unité.

KEYWORDS: Concaténation cost, corpus-based TTS, Unit Selection.

1 Introduction

Au cours des dernières années, la recherche en synthèse de la parole à partir du texte s'est essentiellement portée sur deux techniques. L'approche statistique paramétrique (SPSS pour *Statistical Parametric Speech Synthesis*) est la plus récente et a été l'objet de nombreux travaux universitaires ces dernières années. Elle comprend principalement la synthèse par HMM et plus récemment par

DNN (Black *et al.*, 2007; Yamagishi *et al.*, 2008; Hashimoto *et al.*, 2015). Cette méthode offre un contrôle avancé sur le signal et produit une synthèse très intelligible, mais la voix générée manque de naturel. La méthode historique, la Synthèse Par Corpus (SPC), est un raffinement de la synthèse par concaténation (Sagisaka, 1988; Hunt & Black, 1996; Taylor *et al.*, 1998; Breen & Jackson, 1998; Clark *et al.*, 2007). La SPC permet la création de synthèse de haute qualité, dont le naturel et la qualité prosodique restent inégalés par les autres méthodes grâce à l'utilisation de parole naturelle pour réaliser la synthèse. La plupart des systèmes industriels fonctionnent grâce à cette méthode qui a cependant des inconvénients, telle la difficulté à contrôler la prosodie et le risque d'artefacts de concaténation pénalisant l'intelligibilité.

Cette méthode fait intervenir la notion d'unité, laquelle est une liste de segments contigüs (des diphones généralement) dans un corpus de parole correspondant à une partie de la séquence cible de segments à synthétiser. Afin de discriminer les segments provenant du corpus qui correspondent aux besoins exprimés par l'intermédiaire de la séquence cible, la méthode habituelle est de classer les unités en évaluant le degré de ressemblance avec la séquence cible (coût cible) et le risque de créer un artefact lors de la concaténation des unités (coût de concaténation) via des fonctions des coûts. Le coût de concaténation repose principalement sur des caractéristiques acoustiques (MFCC, F0) (Stylianou & Syrdal, 2001; Tihelka *et al.*, 2014) pour évaluer le niveau de ressemblance spectrale entre deux stimuli vocaux sur et autour du point de concaténation. Ces coûts de concaténation sont toutefois loin d'être parfaits et de nombreux artefacts apparaissent à la fois dans les systèmes commerciaux et de recherche, même après un traitement post-concaténation. Plusieurs analyses ont montré que ces artefacts se produisent plus souvent sur certains phonèmes que sur d'autres (Yi, 1998; Cadic *et al.*, 2009). La concaténation sur une voyelle est par exemple plus risquée que sur une fricative sourde. Cette observation est à l'origine d'une méthode de construction de script d'enregistrement dans (Cadic *et al.*, 2009) où la couverture de «sandwichs vocaliques» vise à favoriser les concaténations sur des diphonèmes jugés peu risqués. Dans cet article, nous proposons d'intégrer ces contraintes directement dans la fonction de coût, sans l'aide d'un corpus construit avec des sandwichs vocaliques. Nous intégrons ainsi une pénalité en fonction de la classe de phonèmes dans la fonction de coût lors de la sélection d'unité. Deux versions sont proposées : d'abord en utilisant une pénalité fixe puis une fonction floue visant à rendre la pénalisation des unités plus flexible. La version floue a été utilisée dans notre système pour le Blizzard Challenge 2015 (Alain *et al.*, 2015), bien qu'elle n'aie pas été évaluée à l'époque.

L'article est organisé comme suit. Dans la section 2, le système de synthèse est présenté. La section 3 propose l'utilisation de contraintes phonologiques et l'introduction d'une pénalité dans la fonction de coût, de manière à obtenir un meilleur classement des chemins lors de la sélection. La section 4, décrit le corpus utilisé pour l'évaluation de cette nouvelle méthode. Enfin la section 5 présente les résultats des expériences menées pour évaluer les approches proposées de manière objective et perceptive.

2 Le système de synthèse de l'IRISA

Le système de synthèse de l'IRISA (Guennec & Lolive, 2014), utilisé pour les expériences présentées dans ce document, est fondé sur une approche de type sélection d'unité, réalisée via un algorithme optimal de recherche de plus court chemin dans un graphe (ici un algorithme A*). Dans les systèmes

de SPC, la fonction optimisée est habituellement écrite comme suit (Hunt & Black, 1996) :

$$U^* = \underset{U=u_1,\ldots,u_N}{\arg\min} \left(W_{tc} \sum_{n=1}^{N} C_t(u_n) + W_{cc} \sum_{n=2}^{N} C_c(u_{n-1}, u_n)\right) \quad (1)$$

où U^* est la meilleure séquence d'unités selon la fonction de coût et u_n est l'unité candidate que l'on essaie de faire correspondre à la n$^{\text{ème}}$ unité cible dans la séquence candidate U. $C_t(u_n)$ est le coût cible et $C_c(u_{n-1}, u_n)$ est le coût de concaténation. W_{tc} et W_{cc} sont les pondérations associées aux deux sous-coûts (Alías *et al.*, 2011). Ces poids sont calculés à l'aide des distributions de coût observées dans le corpus et visent à compenser les ordres de grandeur des sous-coûts comme dans (Blouin *et al.*, 2002). Le coût de concaténation est composé de distances euclidiennes sur les MFCC (hors coefficients dérivés Δ et $\Delta\Delta$), l'amplitude et la F0 :

$$C_c(u_{n-1}, u_n) = C_{mfcc}(u_{n-1}, u_n) + C_{amp}(u_{n-1}, u_n) + C_{F0}(u_{n-1}, u_n) \quad (2)$$

où $C_{mfcc}(u_{n-1}, u_n)$, $C_{amp}(u_{n-1}, u_n)$ et $C_{F0}(u_{n-1}, u_n)$ sont les trois sous-coûts de MFCC, amplitude et F0.

Dans cet article, le coût cible est mis à 0 dans l'équation 1. À la place, nous filtrons les unités candidates du corpus, en n'incluant dans le graphe que celles correspondant à un ensemble de caractéristiques linguistiques et phonétiques, que nous appelons les filtres de présélection (Donovan, 2001). Afin d'obtenir un nombre suffisant d'unités candidates u_n pour chaque unité cible t_n, que nous notons MIN_u (10 au minimum dans ce travail), les contraintes liées aux filtres peuvent être temporairement relâchées. Plus formellement, on considère que l'on dispose d'un n-uplet de J filtres modélisés par des fonctions indicatrices $f_j(u_n, t_n)$ ($j \in [0\,;J]$) valant 1 si u_n respecte la condition posée par le filtre j pour le diphone cible t_n et 0 sinon. Considérons l'ensemble des unités satisfaisant les I premiers filtres pour le diphone cible t_n :

$$O(I_n, t_n) = \left\{ u_n / \prod_{i=1}^{I_n \leq J} f_i(u_n, t_n) = 1 \right\}. \quad (3)$$

L'étape de présélection vise à rechercher, pour chaque diphone cible t_n, l'ensemble $O(I_n, t_n)$ de noeuds candidats, à intégrer dans le graphe de sélection, pour lequel I_n est maximal :

$$I_n = \arg\min card(O(I_n, t_n)) \geq MIN_u. \quad (4)$$

En conséquence, le relâchement des filtres est effectué en partant du dernier. Ainsi, l'ordre des filtres utilisés peut avoir un impact important lors de la sélection.

La principale raison de ne pas intégrer ces filtres à la fonction de coût elle-même est de réduire la taille du graphe d'unités candidates (donc de réduire le temps de sélection). Cependant, il ne faut pas perdre de vue le fait qu'ils font partie intégrante du coût de sélection. En effet, les filtres constituent un ensemble de fonctions binaires de coûts cible se fondant sur l'hypothèse suivante : si une unité ne respecte pas l'ensemble des filtres actifs, elle ne peut pas être utilisée pour la sélection. Les filtres de présélection utilisés dans ce travail (choisis empiriquement) sont les suivants :

1. Label du segment associé, diphonème ou autre (ne peut être relâché).
2. Est-ce un *Non Speech Sound* (ne peut être relâché) ?
3. Le phone est-il dans la dernière syllabe du groupe de souffle ?

4. Le phone est-il dans la dernière syllabe de la phrase ?
5. La syllabe courante est-elle en fin de mot ?

On pourrait faire valoir que plus de filtres permettrait une meilleure sélection, mais par expérience plus de raffinement dans les filtres ne s'avère pas donner de meilleurs résultats. En effet, la meilleure unité est essentiellement un compromis entre un bon coût cible et un bon coût de concaténation, ce qui implique que chaque coût doit disposer d'un panel de choix suffisant.

3 Proposition

3.1 Spécification du coût de concaténation

L'analyse des phrases contenant des artefacts de concaténation montre que certains phonèmes, en particulier les voyelles et semi-voyelles, sont plus susceptibles d'engendrer des ruptures spectrales que d'autres (occlusives et fricatives par exemple) (Yi, 1998). Les phonèmes voisés, présentant une énergie acoustique élevée ou fortement dépendants du contexte sont généralement soumis à plus de distorsions. Sur la base de cette constatation, (Cadic *et al.*, 2009) propose un critère de couverture de corpus visant à optimiser la couverture d'unités dites « sandwich ». Un sandwich vocalique est une séquence de phonèmes où un ou plusieurs noyaux syllabiques sont entourés de deux phonèmes considérés comme peu susceptibles de provoquer des artefacts lors de la concaténation (*i.e.* résistantes aux artefacts de concaténation). En ce qui concerne les coûts de concaténation, l'utilisation de la classe phonétique pour contraindre les phonèmes considérés comme problématiques pour la concaténation n'est pas une idée nouvelle (Donovan, 2001; Yi, 1998). Cependant, dans ces travaux, les coûts sont trop contraignants, essayant de trouver une unité parfaite qui n'existe que rarement. En outre, généralement, quand cette unité n'est pas trouvée, aucune unité moins ambitieuse n'est recherchée.

3.2 Une pénalité floue fondée phonologiquement

Dans notre approche, nous définissons deux méthodes de pénalisation fondées sur trois groupes phonétiques :

V (voyelle) : Voyelles, sur lesquelles une concaténation est difficilement acceptable.

A (acceptable) : Semi-voyelles, liquides, nasales, fricatives voisées et schwa. Ces phonèmes sont vus comme des points de concaténation acceptables, du moins s'il n'y a pas de meilleur choix, mais restent dangereux.

R (résistant) : les phonèmes restants (consonnes non voisées, plosives voisées), considérés comme de bons points de concaténation.

Un point clé de la méthode est de limiter le nombre de classes (seulement 3 sous-ensembles de phonèmes ici), ceci afin de ne pas ajouter trop de contraintes dans la fonction de coût. Le but de la pénalité n'est pas d'agir comme un coût à part entière, mais simplement d'introduire des connaissances qui ne sont pas capturées par le coût de concaténation pour affiner le classement des unités. Il convient de noter que les classes proposées ici sont basées sur le bon sens et il peut être nécessaire de les adapter en fonction de la langue, voire d'effectuer une étude plus poussée.

La première méthode pour appliquer la pénalité, appelée *pho-class*, consiste à attribuer une pénalité fixe $p(v)$ dépendante de la classe du phonème qui débute l'unité v : 0 pour les phonèmes de R, une pénalité légèrement supérieure à la valeur la plus élevée de C_c observée dans le corpus pour tous les phonèmes dans A. Dans notre système, pour des coûts de concaténations dans l'intervalle $[0\,;6]$, la pénalité vaut 7 pour la classe A. On attribue une pénalité bien plus importante aux voyelles (V), assez grande pour empêcher toute compensation par d'autres coûts dans la séquence candidate (une pénalité de valeur 100 est appliquée dans notre cas). Une concaténation sur les voyelles ne peut donc intervenir que si l'on n'a pas d'autre choix. Dans ce cas, un nouveau coût de concaténation C'_c est formulé comme suit :

$$C'_c(u,v) = C_c(u,v) + K(u,v) \tag{5}$$

avec $K(u,v) = p(v)$.

La deuxième méthode, appelée *fuzzy-pho-class*, consiste à moduler la pénalité dans certains cas. Ainsi, nous introduisons une fonction floue de pondération multipliant chaque pénalité par un poids compris entre 0 et 1, comme le montre la figure 1. Elle décrit le degré de satisfaction de l'unité candidate à l'égard de la qualité de concaténation. En faisant l'hypothèse que les distributions des coûts de MFCC, d'amplitude de F0 définies dans (2) suivent des lois normales, nous définissons deux seuils pour chaque sous-coût. Ces distributions sont estimées à l'aide du corpus de parole en calculant le sous-coût de concaténation pour la F0, l'amplitude et les MFCC en utilisant toutes les unités présentes dans le corpus. Par exemple, les deux seuils T^1_{F0} et T^2_{F0} pour le sous-coût de F0 peuvent être définis comme suit :

$$\begin{aligned} T^1_{F0} &= \mu_{F0} - \sigma_{F0} \\ T^2_{F0} &= \mu_{F0} + \sigma_{F0} \end{aligned} \tag{6}$$

où μ_{F0} et σ_{F0} désignent l'espérance et l'écart-type de $C_{F0}(u,v)$. Formellement, la fonction floue pour le sous-coût de F0 est définie comme suit :

$$f_{F0}(u,v) = \begin{cases} 0 & \text{si } C_c(u,v) < T^1_{F0}, \\ 1 & \text{si } C_c(u,v) > T^2_{F0}, \\ 1 - \frac{(T^2_{F0} - C_c(u,v))}{(T^2_{F0} - T^1_{F0})} & \text{sinon.} \end{cases} \tag{7}$$

Le choix de cet intervalle de tolérance est motivé par l'observation de la répartition des coûts réels sur le corpus de voix. Pour être complet, le choix des seuils devrait être différencié selon le type de sous-coût et optimisé séparément (ce qui n'est pas le cas ici). Enfin, la pénalité est modifiée de la façon suivante :

$$K(u,v) = (f_{mfcc}(u,v) + f_{amp}(u,v) + f_{F0}(u,v)) * p(v)$$

où $f_{mfcc}(u,v)$, $f_{amp}(u,v)$ et $f_{F0}(u,v)$ correspondent aux fonctions floues de la forme décrite dans la figure 1 pour les MFCC, l'amplitude et la F0 respectivement (la figure prend l'exemple de la F0, mais les autres fonctions sont identiques). Avec ces fonctions, l'idée principale est de diminuer la pénalité lorsque l'unité a une valeur de sous-coût de concaténation qui est statistiquement parmi les meilleures. Si le coût de concaténation est au-dessus du seuil le plus élevé alors la pénalité complète doit être appliquée car l'unité considérée est alors parmi les pires possibles. Entre les deux seuils, la pénalité est progressivement augmentée au fur et à mesure que le coût de concaténation augmente.

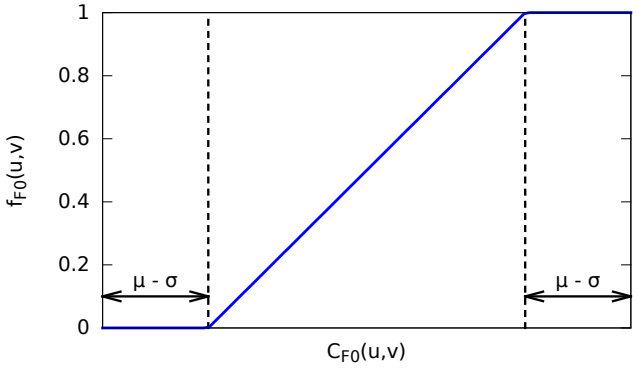


FIGURE 1 – La fonction floue $f_{F0}(u, v)$ (ici pour la F0) sur la distribution des sous-coûts $C_{F0}(u, v)$. Le poids 0 (resp. 1) est accordé aux unités qui ont un coût de concaténation parmi les 15 % les plus faibles (resp. les plus élevés). Entre ces seuils, le poids augmente de manière linéaire.

4 Description du corpus

Pour réaliser les expériences, deux corpus de parole ont été utilisés. Le premier est un corpus extrait d'un livre audio expressif, nommé ici *Audiobook*. Il contient 10h45 de parole et est échantillonné à 44,1 kHz avec un encodage sans perte en mono-canal. Le locuteur est un homme et la F0 moyenne des segments voisés est faible, à seulement 87Hz dans l'ensemble du corpus. Les données ont été automatiquement annotées en utilisant le procédé décrit dans (Boeffard *et al.*, 2012) et en utilisant le toolkit ROOTS (Chevelu *et al.*, 2014). Le corpus est composé de 3339 énoncés, avec 388251 phonèmes et 31491 NSS (*Non Speech Sound*). Sa couverture de diphonème est incomplète à 78 %, mais seuls des diphonèmes très rares/irréalisables en français en sont exclus.

Le second corpus utilisé pour nos expériences est ici nommé *IVS*. Il a été enregistré à des fins de synthèse de parole avec pour but de l'intégrer à un système vocal interactif. Le script d'enregistrement a été construit spécifiquement pour couvrir tous les diphonèmes présents en français. Il comprend également un ensemble de mots très usités dans le domaine des télécommunications. Il dispose d'une voix féminine échantillonnée à 16kHz (codage sans perte, 1 canal) avec une F0 moyenne sur les segments voisés assez basse à 163Hz. Le corpus est composé de 7662 énoncés, 239260 phonèmes et 20424 NSS pour 7h05' de parole. Le corpus de test utilisé pour les expériences consiste en 100 phrases issues de styles très différents. Ce corpus est distinct d'*IVS* et d'*Audiobook*. La langue des corpus utilisés est le français.

5 Expériences

Premièrement, une analyse du comportement de nos trois méthodes (*baseline*, *pho-class* et *fuzzy-pho-class*) en termes de coût de concaténation est effectuée. Les unités résistantes (classe R) sont considérées à part. Ensuite, l'évaluation subjective que nous avons menée pour valider l'approche proposée est présentée avec ses résultats.

IVS	Unités résistantes (R)		Unités non résistantes (A, V)		Toutes les unités	
	μ (*std*)	Nb.	μ (std)	Nb.	μ (std)	Nb.
baseline	2.90 (0.69)	582	3.14 (0.70)	1249	3.06 (0.71)	1831
pho-class	3.28 (0.92)	1025	3.35 (0.88)	813	3.31 (0.90)	1838
fuzzy-pho-class	3.35 (0.92)	**1095**	**2.58** (0.42)	1169	**2.95** (0.80)	2264
Audiobook	Unités résistantes (R)		Unités non résistantes (A, V)		Toutes les unités	
	μ (std)	Nb.	μ (std)	Nb.	μ (std)	Nb.
baseline	2.44 (0.52)	606	2.90 (0.60)	1057	2.74 (0.61)	1663
pho-class	2.65 (0.71)	865	3.14 (0.78)	785	2.88 (0.78)	1650
fuzzy-pho-class	2.65 (0.64)	**907**	**2.47** (0.38)	1139	**2.55** (0.52)	2046

TABLE 1 – Coûts de concaténation sans pénalité suivant les trois stratégies (pénalités soustraites *a posteriori*). R, A et V désignent les classes introduites section 3.2, et Nb. le nombre de concaténations.

5.1 Analyse des coûts de concaténation

Premièrement, nous avons étudié l'évolution des coûts en utilisant les trois systèmes et en comparant (1) les coûts moyens de concaténation des unités résistantes seulement (classe R), (2) les unités non-résistantes seulement (classes A, V) et (3) les deux à la fois, à chaque fois en excluant de nos statistiques les phonèmes contigüs. Tous ces résultats sont présentés dans la table 1. Comme on peut le voir, le système *baseline* a des coûts moins élevés que le système *pho-class*, à la fois pour les unités résistantes et non résistantes. L'explication de ce fait est que *pho-class*, en pénalisant les unités non-resistantes, favorise les unités résistantes même si leur coût de concaténation est plus important. Le nombre de concaténations effectuées sur les unités résistantes (1025 pour *IVS*) est significativement plus élevé en comparaison du système *baseline* (582 pour *IVS*). En ce qui concerne *fuzzy-pho-class*, les résultats en terme de nombre de concaténations sont plus nuancés. En effet, étant donné que les faibles coûts de concaténation sur les unités non résistantes sont moins pénalisés, le système *fuzzy-pho-class* obtient le coût le plus faible pour les unités non résistantes. L'introduction de pénalités variables permet d'évaluer les unités avec plus de précision (et plus de pertinence) qu'avec *pho-class*, pour lequel toutes les unités pénalisées le sont de manière équivalente. En contrepartie, le nombre de concaténations augmente globalement pour le système *fuzzy-pho-class*, ce qui n'est pas un problème étant donné que celles-ci sont mieux choisies. Il est intéressant de mentionner que ces résultats se retrouvent sur les deux voix. On peut alors les considérer comme – raisonnablement – indépendants du type de voix.

Pour résumer, la pénalité semble se comporter comme prévu, permettant de favoriser des séquences d'unités avec un moindre coût sur les unités sensibles tout en maximisant les concaténations sur les unités jugées résistantes.

5.2 Évaluation subjective

Pour évaluer les améliorations apportées par les deux méthodes proposées, nous avons conduit deux évaluations subjectives de type MUSHRA impliquant respectivement 8 et 9 auditeurs pour les voix *IVS* et *Audiobook*. Chaque test comprenait 15 étapes, les testeurs écoutant pour chacune 3 échantillons de parole synthétisée de la même phrase, une pour chacun des trois systèmes *baseline*, *pho-class* et

	IVS	*Audiobook*
baseline	51.68 ± 3.51	49.83 ± 3.48
pho-class	56.72 ± 3.59	50.38 ± 3.48
fuzzy-pho-class	$\mathbf{57.34} \pm 3.50$	$\mathbf{53.06} \pm 3.42$

TABLE 2 – Résultats des tests d'écoute avec intervalles de confiance à 95%. L'approche *fuzzy-pho-class* obtient le meilleur score, suivi par *pho-class* puis finalement par *baseline*.

fuzzy-pho-class. Les testeurs notaient ensuite la qualité globale de chaque échantillon sur une échelle de 0 (mauvais) à 100 (excellent). Les conditions de test et le choix des échantillons sont conformes aux recommandations de l'UIT-T.

Les résultats de ces tests sont présentés dans la table 2. Pour les deux voix, l'approche *fuzzy-pho-class* obtient les meilleurs résultats, suivie du système *pho-class* avec des résultats intermédiaires. Il est intéressant de noter que dans le cas de la voix *IVS*, ces deux systèmes obtiennent des scores similaires. À l'inverse, la différence est plus importante pour la voix *Audiobook*. L'explication de ce phénomène réside certainement dans le fait que dans le cas d'une voix neutre (*IVS*), la faible variabilité des unités renforce les résultats de l'approche *pho-class*. À l'inverse, dans le cas d'une voix plus expressive – comme c'est le cas pour *Audiobook*– la variabilité des unités est bien plus importante et les contraintes sur les phonèmes à concaténer sont bien plus fortes. Dans ce cas, *fuzzy-pho-class* est plus efficace grâce à la flexibilité de l'approche floue, bien plus adaptable que la pénalité fixe octroyée par la méthode *pho-class*.

6 Conclusion

Dans cet article, nous avons proposé une nouvelle fonction de coût de concaténation introduisant une pénalité sur la base de contraintes phonologiques. Une seconde approche, nuançant cette pénalité en fonction de la répartition des coûts via une fonction d'appartenance floue, a également été présentée. La pénalité permet d'éviter des artefacts lors de la synthèse. La version floue permet en outre de garder suffisamment de variabilité lors de la sélection. Les expériences subjectives que nous avons menées montrent une meilleure performance de la version floue à la fois sur une voix neutre et une voix expressive. On montre ainsi que le coût de concaténation ne saisit pas toute l'information perceptive et que l'ajout de certaines préférences sur le type d'unités à concaténer améliore la qualité de la parole synthétisée. Des modèles flous plus avancés peuvent maintenant être étudiés, de manière à améliorer encore la méthode. Il est également nécessaire de mener d'autres travaux sur la classification des phonèmes dans les ensembles R, A et V. En effet, la classification sur laquelle ils reposent devrait être comparée à d'autres, peut-être plus adéquates. Par exemple, les liquides et les semi-voyelles se montrant souvent problématiques, on pourrait considérer leur ajout dans la classe V. Plus de classes pourraient également être ajoutées, en apportant toutefois un soin particulier à ne pas se montrer trop contraignant, trop de contraintes dégradant généralement la qualité de la synthèse. Nous souhaitons également mener une étude sur la dépendance de ces classes à la langue. En outre, l'efficacité de l'approche pourrait être évaluée sur des corpus construits à l'aide d'un script d'enregistrement optimisant le taux de couverture en sandwichs vocaliques (suivant la méthodologie de (Cadic *et al.*, 2009)).

Références

ALAIN P., CHEVELU J., GUENNEC D., LECORVÉ G. & LOLIVE D. (2015). The irisa text-to-speech system for the blizzard challenge 2015. In *The Blizzard Challenge*.

ALÍAS F., FORMIGA L. & LLORÁ X. (2011). Efficient and reliable perceptual weight tuning for unit-selection text-to-speech synthesis based on active interactive genetic algorithms : A proof-of-concept. *Speech Communication*, **53**(5), 786–800.

BLACK A. W., ZEN H. & TOKUDA K. (2007). Statistical Parametric Speech Synthesis. *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, **4**.

BLOUIN C., ROSEC O., BAGSHAW P. & D'ALESSANDRO C. (2002). Concatenation cost calculation and optimisation for unit selection in TTS. In *IEEE Workshop on Speech Synthesis*.

BOEFFARD O., CHARONNAT L., MAGUER S. L., LOLIVE D. & VIDAL G. (2012). Towards Fully Automatic Annotation of Audio Books for TTS. In *Proc. of LREC*, p. 975–980.

BREEN A. P. & JACKSON P. (1998). Non-uniform unit selection and the similarity metric within BT's Laureate TTS system. In *Proc. of the ESCA Workshop on Speech Synthesis*, p. 373–376.

CADIC D., BOIDIN C. & D'ALESSANDRO C. (2009). Vocalic sandwich, a unit designed for unit selection TTS. In *Tenth Conference of ISCA*, p. 2079–2082.

CHEVELU J., LECORVÉ G. & LOLIVE D. (2014). ROOTS : a toolkit for easy, fast and consistent processing of large sequential annotated data collections. In *Proc. of LREC*, p. 619–626.

CLARK R. A., RICHMOND K. & KING S. (2007). Multisyn : Open-domain unit selection for the Festival speech synthesis system. *Speech Communication*, **49**(4), 317–330.

DONOVAN R. E. (2001). A new distance measure for costing spectral discontinuities in concatenative speech synthesizers. In *ITRW*.

GUENNEC D. & LOLIVE D. (2014). Unit Selection Cost Function Exploration Using an A* based Text-to-Speech System. In *Proc. of TSD*, p. 432–440.

HASHIMOTO K., OURA K., NANKAKU Y. & TOKUDA K. (2015). The Effect Of Neural Networks In Statistical Parametric Speech Synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, p. 4455–4459, Melbourne.

HUNT A. J. & BLACK A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. of ICASSP*, volume 1, p. 373–376.

SAGISAKA Y. (1988). Speech synthesis by rule using an optimal selection of non-uniform synthesis units. In *Proc. of ICASSP*, p. 679–682.

STYLIANOU Y. & SYRDAL A. K. (2001). Perceptual and objective detection of discontinuities in concatenative speech synthesis. *Proc. of ICASSP*, **2**, 837–840.

TAYLOR P., BLACK A. W. & CALEY R. (1998). The architecture of the Festival speech synthesis system. In *Proc. of the ESCA Workshop in Speech Synthesis*, p. 147–151.

TIHELKA D., MATOUŠEK J. & HANZLÍČEK Z. (2014). Modelling F0 Dynamics in Unit Selection Based Speech Synthesis. In *International Conference on Text, Speech and Dialogue*, p. 457–464.

YAMAGISHI J., LING Z. & KING S. (2008). Robustness of HMM-based speech synthesis. In *Ninth Annual Conference of the International Speech Communication Association*.

YI J. (1998). *Natural-sounding speech synthesis using variable-length units*. Rapport interne, Massachusetts Institute of Technology.

Perception audio-visuelle de séquences VCV produites par des personnes porteuses de Trisomie 21 : une étude préliminaire

Alexandre Hennequin[1], Amélie Rochet-Capellan[2] & Marion Dohen[1]
(1) Univ. Grenoble Alpes, GIPSA-Lab, F-38000 Grenoble, France
(2) CNRS, GIPSA-Lab, F-38000 Grenoble, France
{alexandre.hennequin, amelie.rochet-capellan, marion.dohen}@gipsa-lab.grenoble-inp.fr

RESUME

La parole des personnes avec trisomie 21 (T21) présente une altération systématique de l'intelligibilité qui n'a été quantifiée qu'auditivement. Or la modalité visuelle pourrait améliorer l'intelligibilité comme c'est le cas pour les personnes « ordinaires ». Cette étude compare la manière dont 24 participants ordinaires perçoivent des séquences VCV voyelle-consonne-voyelle) produites par quatre adultes (2 avec T21 et 2 ordinaires) et présentées dans le bruit en modalités auditive, visuelle et audiovisuelle. Les résultats confirment la perte d'intelligibilité en modalité auditive dans le cas de locuteurs porteurs de T21. Pour les deux locuteurs impliqués, l'intelligibilité visuelle est néanmoins équivalente à celle des deux locuteurs ordinaires et compensent le déficit d'intelligibilité auditive. Ces résultats suggèrent l'apport de la modalité visuelle vers une meilleure intelligibilité des personnes porteuses de T21.

ABSTRACT

Auditory-visual Perception of VCVs Produced by People with Down Syndrome: a Preliminary Study

The speech of people with Down Syndrome (DS) is systematically altered resulting in an intelligibility loss. This was quantified only auditorily. The visual modality could actually improve intelligibility, as is the case for "ordinary" people. The present study compares the way 24 ordinary participants perceive VCV sequences (vowel-consonant-vowel) produced by four adults (2 with DS and 2 ordinary) and presented in noise in three modalities: auditory, auditory-visual and visual. The results confirm an intelligibility loss in the auditory modality for speakers with DS. However, for the two speakers involved in this study, visual intelligibility is equivalent to that of the ordinary speakers and compensates for the auditory intelligibility loss. These results put forward the importance of integrating multimodality to improve the intelligibility of people with DS.

MOTS-CLES : Parole, Multimodalité, Perception, Trisomie 21, Apport visuel.

KEYWORDS: Speech, Multimodality, Perception, Down Syndrome, Visual input.

1 Introduction

La Trisomie 21 (T21) est une anomalie génétique très fréquente, présente dans toutes les sociétés et causée par la présence d'un chromosome 21 surnuméraire dans le génotype. Cette anomalie induit des troubles anatomiques, physiologiques et cognitifs. Elle est la première cause génétique de déficience intellectuelle (Katz & Lazcano-Ponce, 2008). Le trouble de production de la parole est systématique et n'est pas seulement imputable à la déficience intellectuelle (Kumin, 2006 ; Bunton *et al*., 2007 ; Kent & Vorperian, 2013). Les compétences intellectuelles d'un individu étant souvent inférées de sa capacité à s'exprimer, le trouble de la parole des personnes avec T21 est un enjeu de prise en charge central pour améliorer leur intégration sociale. Or peu d'études ont quantifié l'intelligibilité des personnes avec T21 et la manière dont les personnes « ordinaires »[1] non familiarisées perçoivent leur parole. De plus, les études sur la perception de la parole produite par des locuteurs tout-venant montre que la vision du visage du locuteur améliore la perception de sa parole, notamment en milieu bruité (e.g. Grant & Seitz, 2000). Dans ce cadre, notre objectif est d'évaluer si la vision aide à la perception de la parole produite par des locuteurs avec T21.

Kent et Vorperian (2013) ont publié une revue des travaux de recherche sur la production de la parole par les locuteurs avec T21 depuis les années 1950 selon 4 axes : la voix, l'articulation, la fluence/prosodie et l'intelligibilité. Cette revue montre d'abord que l'intérêt pour la recherche sur la parole des personnes avec T21 a récemment augmenté surtout concernant les aspects articulatoires. Les résultats des études menées sont globalement mitigés et parfois contradictoires du fait du nombre limité de participants et de l'utilisation de méthodologies très variées. Par exemple, bien que la fréquence fondamentale (F0) soit perçue comme étant plus faible chez les personnes avec T21, les résultats d'études acoustiques suggèrent plutôt des valeurs de F0 plus importantes chez ces personnes. La qualité vocale des personnes avec T21 est souvent décrite comme rauque mais cette impression n'est pas quantifiée. Les personnes avec T21 font beaucoup d'erreurs articulatoires et/ou phonologiques dans la production de mots qui rappellent la parole dysarthrique (Bunton *et al*., 2007 ; Kumin, 2006). La littérature rend aussi compte de dysfluences et de différences prosodiques, le bégaiement est notamment fréquent. Plusieurs études, se basant principalement sur des jugements perceptifs et des questionnaires aux familles, rapportent des problèmes d'intelligibilité (Kumin, 2006, 2012 ; Bunton *et al*., 2007). On note de plus la présence de fortes idiosyncrasies. Par exemple, des scores similaires à un test d'intelligibilité de mots peuvent être associés à des profils d'erreurs différents (Bunton *et al*., 2007). Enfin, le trouble de la parole des personnes avec T21 s'observe dès un très jeune âge, avec des différences de développement tels qu'un retard observable sur le babillage canonique ou la production plus fréquente de sons n'étant pas de la parole.

Toujours selon Kent et Vorperian (2013, voir aussi Kumin, 2012), les difficultés de parole décrites ci-dessus ont des origines très variées : problèmes de contrôle moteur, d'audition, de retours somatosensoriels dans la cavité orale, déficience cognitive (avec notamment un déficit du traitement de l'information auditive sérielle), anomalies physiologiques et anatomiques du conduit vocal. La cavité orale des personnes avec T21 est notamment plus petite donnant l'impression d'une macroglossie (*i.e*. une langue anormalement volumineuse) alors que le pharynx possède des caractéristiques de volume et de taille usuelles. La dentition et le palais sont aussi affectés dans la majorité des cas. Ces anomalies ont des conséquences sur la précision de positionnement des

[1] Le terme « ordinaires » est utilisé ici pour « tout-venant » en accord avec la terminologie préconisée par un des organismes financeurs du project : la FIRAH.

articulateurs dans la production de la parole. De plus, la pression intra orale et l'activation musculaire lors de la production de la parole sont supérieures à celles des personnes « ordinaires ». Ces observations suggèrent que la parole demande un effort moteur particulièrement important aux personnes avec T21. Elles s'accordent avec les travaux sur les mouvements des membres qui suggèrent des seuils d'activation musculaire plus hauts que chez les personnes tout-venant, liés à l'observation d'une hypotonie générale au repos (Latash *et al*., 2008).

Concernant les aspects perceptifs, en regard des spécificités physiologiques et anatomiques décrites ci-dessus, le déficit d'intelligibilité de la parole des personnes avec T21 a été essentiellement décrit dans la modalité auditive. On peut dès lors s'interroger sur le rôle de la modalité visuelle dans la perception de la parole des personnes avec T21. Cette modalité est-elle moins touchée que la modalité auditive ? Peut-elle rendre la parole plus intelligible qu'en modalité auditive seule ? Voir son interlocuteur aide à mieux percevoir et détecter sa parole notamment lorsque celle-ci est perturbée comme en milieu bruité (pour une revue, voir Dohen, 2009). Les informations auditives et visuelles sont de plus complémentaires et non redondantes. Summerfield (1987) a comparé les confusions auditive et visuelle des consonnes en anglais et montre que le mode d'articulation est plus robuste en auditif alors que c'est le lieu d'articulation en visuel. L'apport visuel aide ainsi à la perception de la parole produite par des locuteurs « ordinaires », mais qu'en est-il de celle produite par des locuteurs avec T21 ?

Hustad et Cahill (2003) se sont intéressés à l'apport de la modalité visuelle dans la perception de phrases ayant une faible prédictibilité sémantique produites par 5 locuteurs avec des dysarthries moyennes à sévères. La modalité visuelle s'est révélée avoir un apport pour un locuteur seulement, souffrant d'une dysarthrie sévère. Keintz *et al*. (2007) ont réalisé une étude similaire avec 8 patients souffrant de la maladie de Parkinson associée à une dysarthrie et des auditeurs expérimentés et non-expérimentés. Leurs résultats montrent une amélioration significative de l'intelligibilité en modalité audiovisuelle par rapport à en audio seul pour les 3 locuteurs ayant les scores d'intelligibilité les plus réduits, équivalente pour les deux groupes d'auditeurs.

Ce travail s'intéresse à l'apport de la modalité visuelle pour la perception de la parole de deux jeunes adultes avec T21 (avec une bonne intelligibilité) par des personnes tout-venant (n'ayant jamais ou qu'occasionnellement interagi avec des personnes avec T21) en comparaison de celle de deux adultes tout-venant de même genre et sexe.

2 Méthodologie

2.1 Participants au test perceptif

24 personnes ont participé à cette étude, toutes de langue maternelle française (12 femmes ; âge : 25,1 (moy) ±3 (e.t.)). Aucune n'a rapporté de problèmes de vision non corrigé, de trouble phonologique ou de la parole. Toutes ont passé un test audiométrique et aucun déficit auditif n'a été constaté. Ils ont été dédommagés pour leur participation par une carte cadeau de 15€.

2.2 Locuteurs et stimuli

Quatre locuteurs de langue maternelle française ont été sélectionnés parmi les locuteurs enregistrés pour une étude précédente (voir Rochet-Capellan & Dohen, 2015) : 2 personnes « ordinaires » (1 homme, 22 ans – 1 femme, 21 ans) et 2 personnes avec T21 (1 h, 21 ans – 1 f, 19 ans). Les locuteurs avec T21 ont été choisis pour leur relativement bonne intelligibilité d'après un

pré-test de perception et les locuteurs « ordinaires » par appariement en âge et en genre. Le pré-test d'intelligibilité a été réalisé en modalité auditive et sans bruit auprès de participants experts sur un ensemble de 7 locuteurs avec T21.

Les stimuli audio-visuels correspondent à 16 séquences de type Voyelle-Consonne-Voyelle (VCV) avec V=[a] et C={[b], [d], [g], [p], [t], [k], [f], [s], [ʃ], [v], [z], [ʒ], [l], [ʁ], [m], [n]}. Chaque VCV était produit 3 fois et nous avons choisi comme stimulus pour le test perceptif la plus claire de ces trois répétitions aux niveaux auditif et visuel. L'enregistrement des stimuli a été réalisé en chambre sourde. Les participants étaient assis, portaient un micro serre-tête (Sennheiser HP4) et étaient filmés avec une caméra numérique HD (Panasonic HC-X920). Ils devaient répéter des séquences VCV qu'ils entendaient via un haut-parleur. L'audio a été échantillonné à 44100 Hz (carte son externe Focusrite Scarlett 6i6). Chaque fichier audio a été normalisé en intensité à 70dB avec Praat puis ajouté à un bruit de type « cocktail party » (BDBRUIT, Zeiliger *et al*., 1994) avec un rapport signal sur bruit de -4dB. Les fichiers audio ont ensuite été resynchronisés aux vidéos (logiciel FFmpeg : https://www.ffmpeg.org, résolution 960x540 pixels) et créés en 3 versions : audio seul (A, avec image d'un haut parleur), vidéo seul (V) et audio vidéo (AV) soit un total de 192 stimuli : 4 locuteurs x 3 conditions x 16 VCV.

2.3 Procédure pour le test perceptif

Les participants au test perceptif étaient assis devant un bureau, à 60 cm d'un écran de 24 pouces, et portaient un micro casque (Audio Technica BPHS1). Le signal acoustique a été numérisé à 48 kHz (carte son externe Focusrite Scarlett 6i6). L'expérience a été programmée en utilisant la Psychophysics Toolbox (http://psychtoolbox.org/, sous Matlab). Le test perceptif était divisé en trois blocs correspondant aux trois modalités (A, V et AV) de 64 stimuli chacun (16 VCVs x 4 locuteurs). L'ordre des blocs était contrebalancé d'un participant à l'autre. L'ordre des stimuli à l'intérieur de chacun des blocs était aléatoire d'un bloc à l'autre et d'un participant à l'autre.

Les participants étaient informés qu'ils allaient entendre, voir, ou voir et entendre une personne prononcer un son de parole deux fois de suite et qu'ils devraient répéter ce qu'ils avaient perçu. La répétition a été utilisée pour s'affranchir d'ambiguïtés de transcription orthographique. Il leur était précisé d'interpréter les séquences comme des sons n'ayant aucun sens mais aucune information n'était fournie sur la structure du son. Après un entraînement avec des stimuli non bruités, un extrait du bruit leur était présenté. Chaque essai avait la structure suivante : la vidéo intégrant deux répétitions du stimulus était jouée au centre de l'écran. Après avoir vu et/ou entendu les deux répétitions, le participant donnait sa réponse orale puis appuyait sur une touche d'un clavier pour passer au stimulus suivant.

2.4 Transcription des réponses et analyses

Les réponses audio fournies par les participants ont été retranscrites selon le code suivant : avantV1-V1-C-V2-aprèsV2. Chaque partie a été transcrite phonétiquement ou codée comme vide (ex : « brata » pour « ata », avantV1='br' – V1='a' – C='t' – V2='a' – aprèsV2=''). Une consonne non perçue était codée par 'h'. Si la réponse était une voyelle unique (ex : 'a' pour 'ata'), elle était codée en V2 (V1='' – C='h' – V2='a'). Une réponse impossible à retranscrire était codée : V1=' ?' – C=' ?' – V2=' ?'. Une réponse correcte désigne le cas où **V1**, **C** et **V2** correspondent à la stimulation et où **avantV1**='' et **aprèsV2**=''. Notre mesure principale est le nombre de réponses correctes. Une analyse plus détaillée des erreurs sur les consonnes et les voyelles a aussi été réalisée. L'analyse des résultats a été faite avec le logiciel R (https://www.r-

project.org/) avec des analyses de la variance (ANOVA). Les comparaisons post-hoc ont été réalisées avec des tests de Student avec correction de Bonferroni.

3 Résultats

L'analyse montre que 44,2% des réponses fournies par les participants étaient correctes et 54,4% comportaient au moins une erreur. Seules 1,4 % des réponses n'ont pas pu être transcrites. Le groupe de locuteurs et la modalité n'ont pas d'effet significatif sur ce pourcentage ($p > 0,1$). On rappelle que pour les résultats obtenus, le niveau de chance est à 6.25% (16 séquences VCV possibles).

3.1 Réponses correctes

La figure 1 présente les pourcentages de réponses correctes en fonction du groupe de locuteurs (Ord vs T21) et de la modalité de présentation (AV vs A vs V). L'ANOVA dont les résultats sont reportés ci-dessous comportait deux facteurs intra-sujets (groupe de locuteurs et modalité) et un facteur inter-sujets (ordre de présentation des modalités).

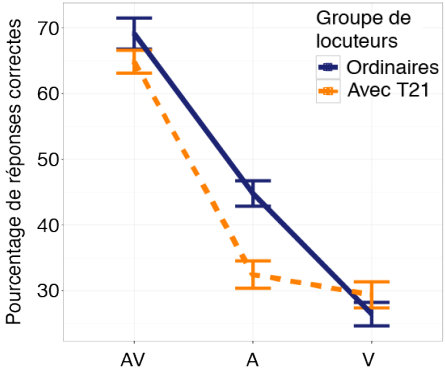


FIGURE 1 : Pourcentage de réponses correctes en fonction du groupe de locuteurs (Ord, T21) et de la modalité

La modalité a un effet significatif sur ce pourcentage ($F(2,36)=263.5$ – $p < 0.001$) : c'est en modalité AV qu'on obtient le plus de réponses correctes suivie des modalités A puis V (A vs AV : $t(23)=-12.8$; $p < 0.001$ – A vs V : $t(23)=5.4$ $p < 0.001$). Les résultats sont globalement meilleurs pour les locuteurs ordinaires que pour les locuteurs porteurs de T21 ($F(1,18)=14.6$ – $p = 0.001$). Ceci dépend cependant de la modalité (groupe de locuteurs * modalité : $F(2,36)=13.6$ – $p < 0.001$) : en modalité A, les locuteurs ordinaires sont significativement mieux perçus que ceux avec T21 ($t(23)=6.7$ – $p < 0.001$) mais ça n'est pas le cas en modalité V ($t(23) = -0.94054$ – $p > 0.9$). En AV, on observe seulement une tendance (non significative après correction pour les comparaisons multiples) à ce que les locuteurs ordinaires soient mieux perçus que ceux porteurs de T21 ($t(23)=2,1$ – $p > 0.1$). L'ordre de passage des modalités a également un effet significatif ($F(5,18)=3,2$ – $p < 0.05$) et interagit avec la modalité ($F(10,36)=3,2$ – $p < 0.01$). L'effet d'ordre s'observe seulement dans la modalité V : les résultats en modalité V sont meilleurs quand elle est passée après la modalité AV plutôt qu'avant.

3.2 Erreurs

Les erreurs ont été classées en trois catégories : insertions avant/après la séquence VCV (avantV1≠'' et/ou aprèsV2≠''), erreurs sur la première/seconde voyelle et erreurs sur la consonne. On notera que ces erreurs ne sont pas exclusives entre elles : il est par exemple possible d'avoir une erreur sur la consonne et une autre sur la ou les voyelles. Les erreurs les plus fréquentes sont celles sur la consonne (70,1% du nombre total d'erreurs) et il y a relativement peu d'erreurs sur V1 et V2 (respectivement, 7,1 et 7%) et d'insertions avant ou après la séquence VCV (7,9% pour les deux).

Réponses sur C – Les réponses sur C ont été classées en catégories : correcte, confusion (avec une autre consonne) et autre (*e.g.,* ajout d'une ou plusieurs consonnes). La figure 2 présente la

répartition des réponses sur C en fonction de la modalité, du groupe de locuteurs et du type de réponse. Les réponses correctes et les confusions représentent 94,5% du total des réponses sur la consonne. L'ANOVA a été réalisée sur les pourcentages d'erreurs avec trois facteurs intra-sujets : groupe de locuteurs, modalité et type d'erreur.

Les confusions avec une autre consonne sont les erreurs les plus fréquentes ($F(1,23)=228,3$ – $p<0,001$) et il y en a significativement plus en modalités A et V qu'en AV ($F(2,46)= 206,4$ – $p <0,001$). On constate qu'il n'y a globalement pas de différence entre les groupes ($F(1,23)=2,4$ $p= 0,1$) bien qu'en modalité A, il y ait plus d'erreurs pour les locuteurs avec T21 que pour les ordinaires (groupe de locuteur * modalité : $F(2,46)=3,6$ – $p<0,05$).

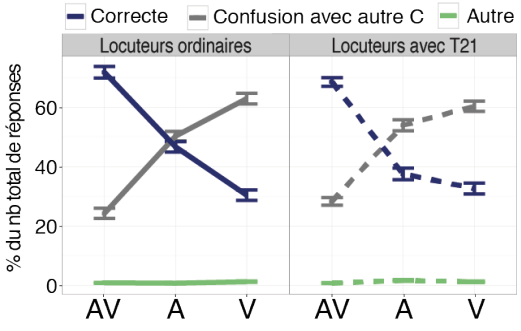


FIGURE 2 : Répartition des réponses sur la consonne en fonction du type de réponse, du groupe de locuteurs (Ord vs T21) et de la modalité (AV, A, V).

Erreurs sur V1 et V2 – Elles ont été classées par catégorie : confusion avec une autre voyelle et voyelle non perçue. La figure 3 fournit les pourcentages d'erreur sur V1 (haut) et sur V2 (bas) en fonction de la modalité, du groupe de locuteurs et du type d'erreur. Les ANOVA réalisées comportent trois facteurs intra-sujets : groupe de locuteurs, modalité et type de réponse.

Erreurs sur V1 – Il y a significativement moins d'erreurs sur V1 en modalité AV qu'en A et V ($F(2,46)=4$ – $p<0,05$). Les erreurs sont significativement plus fréquentes pour les locuteurs avec T21 que chez les ordinaires ($F(1,23)=22,1$ – $p<0,001$) mais seulement en modalité A (groupe de locuteur * modalité : $F(2,46)=11,8$ – $p<0,001$). Il n'y a pas de différence entre les types d'erreurs ($F(1,23)=0,001$ – $p=0,98$). En modalité A, alors que pour les locuteurs ordinaires on observe plus de confusions que de non-perceptions, c'est l'inverse pour les locuteurs avec T21 (groupe de locuteur * modalité * type de réponse : $F(2,46)=6,8$ – $p<0,001$).

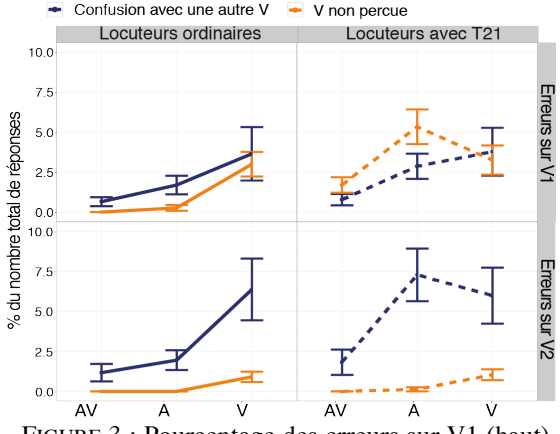


FIGURE 3 : Pourcentage des erreurs sur V1 (haut) et V2 (bas) en fonction du type d'insertion, du groupe de locuteurs (Ord vs T21) et de la modalité (AV, A, V).

Erreurs sur V2 – Il y a significativement moins d'erreurs sur V2 en modalité AV qu'en A et V ($F(2,46)=6,2$ – $p<0,005$). Les erreurs sont significativement plus fréquentes pour les locuteurs avec T21 que pour les ordinaires ($F(1,23)=8$ – $p<0,01$) mais seulement en modalité A (groupe de locuteur * modalité : $F(2,46)=6,6$ – $p<0,005$) et seulement pour les confusions avec une autre voyelle (groupe de locuteur * modalité * type d'erreur : $F(2,46)=5,3$ – $p<0,01$). Les confusions avec une autre voyelle sont significativement plus fréquentes que les non-perceptions ($F(1,23)=8,3$ – $p<0,01$) mais surtout en modalités A et V (modalité * type d'erreur : $F(2,46)=4,1$ – $p=0,03$).

Insertions avant V1 ou après V2 – Elles ont été classées par catégorie : insertion d'une consonne, insertion d'une voyelle et autre (*e.g.*, insertions multiples). La figure 4 fournit les pourcentages d'insertions avant V1 (haut) et après V2 (bas) en fonction de la modalité, du groupe de locuteurs et du type d'insertion. Les ANOVA réalisées comportent trois facteurs intra-sujets : groupe de locuteurs, modalité et type d'insertion.

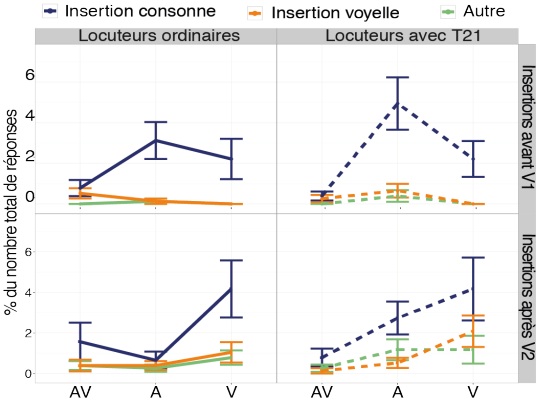


FIGURE 4 : Pourcentage des insertions avant V1 (haut) et après V2 (bas) en fonction du type d'insertion, du groupe de locuteurs (Ord vs T21) et de la modalité (AV, A, V).

Insertions avant V1 – Les insertions avant V1 sont plus fréquentes en A et V qu'en AV, ce type d'erreur étant quasi inexistant en modalité AV (modalité : $F(2,46)=5.8$ – $p < 0.01$). La catégorie d'erreurs la plus fréquente est l'insertion d'une consonne, les autres types d'insertion ne se produisant quasiment pas ($F(2,46)=213.8$ – $p = 0.001$). Il n'y a pas de différence entre les groupes de locuteurs ($F(1,23)=1.8$ – $p = 0.2$). Cependant, les insertions d'une consonne sont plus fréquentes en A qu'en V pour les locuteurs porteurs de T21 mais pas pour les locuteurs ordinaires (groupe de locuteurs * modalité : $F(2,46)=5.8$ – $p < 0.01$).

Insertions après V2 – Il y a plus d'insertions après V2 en modalité V qu'en A et AV ($F(2,46)=4.1$ – $p < 0.05$) et pour les locuteurs porteurs de T21 que pour les locuteurs ordinaires ($F(1,23)=4.8$ – $p < 0.05$). On ne constate pas d'effet du type d'insertion ($F(2,46)=2.9$ – $p > 0.05$). Pour les locuteurs ordinaires, le pourcentage d'insertions après V2 est équivalent en modalités A et AV mais plus important en V alors que pour les locuteurs porteurs de T21, le nombre d'insertions suis un ordre V>A>AV (modalité * groupe de locuteurs : $F(2,46)=3.8$ – $p < 0.05$).

4 Conclusions et discussion

Cette étude compare la perception multimodale, par des participants tout-venant, de séquences Voyelle-Consonne-Voyelle (VCV) produites par des locuteurs avec T21 à celles produites par des locuteurs « ordinaires ». L'objectif était d'évaluer si le fait de voir le locuteur avec T21 parler en plus de l'entendre aide à mieux percevoir sa parole. Comme il a été largement rapporté dans la littérature (cf. Dohen, 2009 pour une revue), les résultats de cette étude montrent que, quel que soit le locuteur (avec T21 ou « ordinaire »), la parole mélangée à du bruit est mieux perçue en modalité AV qu'en modalité A puis V (AV>A>V). Tous groupes de locuteurs confondus, les erreurs de loin les plus fréquentes sont celles sur la consonne qui impliquent une confusion de celle-ci avec une autre consonne : il s'agissait en effet du seul phonème qui variait au cours du test (Voyelle=[a]). Globalement, les pourcentages de bonnes réponses sont faibles en modalités A et V (moins de 50% de bonnes réponses dans les deux cas) semblant suggérer une très faible intelligibilité pour les deux groupes de locuteurs. Ces pourcentages sont cependant bien supérieurs au hasard (6,25%) et sont liés au nombre important de réponses possibles. Nous nous intéressons dans la suite aux résultats dans chacune des modalités.

Modalité A – Les locuteurs « ordinaires » sont significativement mieux perçus que ceux avec T21, ce qui va dans le sens des études rapportant un déficit de l'intelligibilité auditive chez les

personnes avec T21 (Kumin, 2006 ; Bunton *et al*., 2007; Kent & Vorperian, 2013). En dehors du fait que tous les types d'erreurs soient globalement plus fréquents chez les locuteurs avec T21, il y a plusieurs cas où les différences sont plus importantes que la moyenne. Les insertions d'une consonne avant ou après la séquence VCV sont ainsi particulièrement plus fréquentes pour les locuteurs avec T21 que pour les tout-venants. La deuxième voyelle est de plus beaucoup plus souvent confondue avec une autre voyelle pour les locuteurs avec T21. Cela pourrait s'interpréter par une difficulté accrue à séparer la parole des locuteurs avec T21 du bruit ambiant. On rappelle que le bruit ambiant est de type « cocktail party » c'est-à-dire composé d'une multitude d'autres signaux de parole qui peuvent être confondus avec ou assimilés au signal cible à identifier. Notons que cette séparation des sources est d'autant plus complexe en modalité A. On observe de plus que la première voyelle est beaucoup plus souvent non perçue que confondue chez les locuteurs avec T21 alors que c'est l'inverse chez les tout-venant. Ceci suggère que les locuteurs avec T21 ont des difficultés à produire une voyelle intelligible auditivement à l'initialisation d'une séquence, le manque d'intelligibilité étant ici accentué par la présence de bruit.

Modalité AV – Le fait que l'écart d'intelligibilité entre les groupes soit beaucoup plus faible qu'en modalité A (non significatif après corrections pour les comparaisons multiples) suggère que la modalité visuelle permet de compenser au moins en partie, et pour les deux locuteurs avec T21 de cette étude, le déficit d'intelligibilité auditive. En modalité AV, les types d'erreurs ne dépendent pas du groupe de locuteurs sauf pour la première voyelle pour laquelle on observe significativement plus de non-perceptions pour les locuteurs avec T21.

Modalité V – Dans le cadre de cette étude, utilisant des séquences VCV et testant seulement deux locuteurs avec T21, on observe que les locuteurs avec T21 sont aussi intelligibles visuellement que les tout-venants. Notons que lorsque la modalité V est présentée avant AV, les pourcentages de réponses correctes sont moins bons quel que soit le groupe de locuteur. Cet effet d'ordre est probablement lié au fait que les stimuli sont mieux perçus en AV : les participants ont pu mémoriser l'association audio-visuelle d'un stimulus donné, la confrontation au stimulus visuel peut ensuite servir d'amorce à la réponse pour produire la bonne réponse. Cet effet de l'avantage de l'ordre AV-V pour la modalité V n'a, à notre connaissance, pas été clairement mis en évidence par les études antérieures sur les personnes « ordinaires ». Ce résultat, observé pour les deux groupes de locuteurs, suggère que l'inclusion de personnes avec T21 pourrait augmenter l'attention des participants pour l'information visuelle en condition AV.

Cette étude préliminaire suggère un apport de la modalité visuelle pour améliorer l'intelligibilité des personnes avec T21. Les difficultés sur la première voyelle observées en modalité A et qui se maintiennent en AV font écho aux résultats en contrôle moteur suggérant un seuilplus haut d'activation musculaire chez les personnes avec T21 (Latash *et al*., 2008). Cette difficulté à initialiser le mouvement pourrait diminuer l'énergie sur la première voyelle et rendre son exécution plus difficile que la deuxième voyelle. L'inertie de retour au repos pourrait aussi être plus importante, ce qui pourrait contribuer à expliquer les insertions en fin de VCV. Les erreurs sur la consonne sont probablement liées à l'anatomie du conduit vocal et au manque de précision des points d'articulation portés par la langue. La dépendance des résultats aux lieux et modes d'articulation de la consonne reste à évaluer. Il faudra également corroborer ces résultats pour d'autres voyelles. Des travaux antérieurs ayant montré un apport plus important de la modalité visuelle dans la parole dysarthrique seulement pour les dysarthries sévères (Hustad & Cahill, 2003 ; Keintz *et al*., 2007), on peut aussi s'interroger sur l'effet du choix des locuteurs sur nos résultats : est-ce que l'apport du visuel serait encore plus important pour des personnes avec T21 avec une moins bonne intelligibilité ?

Remerciements

Ce travail a reçu le soutien du European Research Council dans le cadre du 7ème Programme de la Communauté Européenne (FP7/2007-2013 Grant Agreement no.339152- "Speech Unit(e)s"). Il s'inscrit de plus dans le cadre du projet « Communiquons Ensemble » subventionné par la Fondation Internationale de la Recherche Appliquée sur le Handicap (FIRAH). Les auteurs remercient les participants, l'Association de Recherche et d'Insertion Sociale des Trisomiques (ARIST) et les professionnels de l'ESAT-SAJ de l'ARIST.

Références

BUNTON, K., LEDDY, M., & MILLER, J. (2007). Phonetic intelligibility testing in adults with Down syndrome. *Down Syndrome Research and Practice*, 12(1), 1–4.

DOHEN, M. (2009). Speech through the ear, the eye, the mouth and the hand. Lecture Notes in *Computer Science* (including Subseries Lecture Notes in *Artificial Intelligence* and Lecture Notes in *Bioinformatics*), 5398 LNAI, 24–39.

GRANT, K. W., & SEITZ, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, 108(3), 1197-1208.

HUSTAD, K. C. (2008). The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *Journal of Speech, Language & Hearing Research*, 51(3), 562–573.

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.

KATZ, G., & LAZCANO-PONCE, E. (2008). Intellectual disability: definition, etiological factors, classification, diagnosis, treatment and prognosis. *salud pública de méxico*, 50, s132-s141.

KEINTZ, C. K., BUNTON, K., & HOIT, J. D. (2007). Influence of visual information on the intelligibility of dysarthric speech. American Journal of Speech-Language Pathology / American Speech-Language-Hearing Association, 16(3), 222–34.

KENT, R. D., VORPERIAN, H. K., KREIMAN, J., & MAASSEN, B. A. M. (2013). Speech Impairment in Down Syndrome: A Review, *Journal of Speech, Language & Hearing Research*, 56(1), 178–210.

KUMIN, L. (2006). Speech intelligibility and childhood verbal apraxia in children with Down syndrome. *Down's Syndrome, Research and Practice*, 10(1), 10–22.

KUMIN, L. (2012). *Early communication skills for children with Down syndrome: A guide for parents and professionals*. États-Unis : Woodbinehouse

LATASH, M., WOOD, L., & ULRICH, D. (2008) *What is currently known about hypotonia, motor skill development, and physical activity in Down syndrome*. Down Syndrome Education Online (http://www.down-syndrome.org/reviews/2074/).

ROCHET-CAPELLAN, A., & DOHEN, M. (2015). Acoustic characterisation of vowel production by young adults with Down syndrome. Actes de *The International Congress of Phonetic Scences*.

SUMMERFIELD, Q. (1987). Comprehensive Account of Audio-Visual Speech Perception. In: Dodd, B., Campbell, R. (eds.), *Hearing by Eye: The Psychology of Lip-reading*, pp. 3--51. Lawrence Erlbaum Associates, Hillsdale, NJ.

ZEILIGER J., S. J.-F. (1994). BDBRUIT, une base de données parole de locuteurs soumis à du bruit. Dans Actes des *10ièmes journées d'Étude sur la Parole* (pp. 287-290).

Perception des consonnes géminées en japonais langue étrangère par des apprenants francophones

Akiko Takemura[1] Takeki Kamiyama[2, 3]
(1) INALCO 65 rue des Grands Moulins, Paris, France
(2) LECSeL, EA1569, Université Paris 8 2 rue de la Liberté, 93526 Saint-Denis, France
(3) LPP UMR7018, CNRS / Paris 3 USPC 19 rue des Bernardins, 75005 Paris, France
akiko.takemura@gmail.com, takeki.kamiyama@univ-paris8.fr

RESUME

Le japonais présente une opposition phonémique entre les obstruantes simples et géminées, qui pose des difficultés aux apprenants non-natifs tant au niveau de la perception que de la production, notamment quand une opposition similaire est absente dans la langue des apprenants. La discrimination perceptive de cette opposition a été étudiée chez 19 apprenants francophones de deux niveaux différents de compétence et chez 6 auditeurs natifs à l'aide d'une expérience AXB avec des non-mots dysyllabiques prononcés par 2 locuteurs natifs du japonais de Tokyo. Les résultats montrent une différence significative entre les apprenants (10,91% d'erreurs en moyenne) et les natifs (3,86% en moyenne). Le taux d'erreurs était plus élevé quand l'accent lexical du mot testé était du type HB (haut-bas) que BH. Les auditeurs natifs ont également montré un taux d'erreur plus élevé pour la fricative /s/, et aussi quand la consonne est entourée des voyelles fermées /i/ et /u/.

ABSTRACT

Perception of geminate consonants in Japanese as a foreign language by French-speaking learners.
Japanese has a phonemic contrast between singleton and geminate obstruants, which causes difficulty to nonnative learners in perception as well as production, especially when the language of the learners does not have a similar contrast. The perceptual discrimination of this contrast was studied with the aid of an AXB task using disyllabic non-words, administered to 19 French-speaking learners of 2 different proficiency levels and 6 native listeners. The results show a significant difference between the learners (mean error rate of 10.91%) and the native listeners (3.86%). The error rate was higher when the lexical accent of the test word was HL (high-low) than LH. Native listeners also showed a higher error rate for the fricative /s/, and also when the consonant was surrounded by the high vowels /i/ and /u/.

MOTS-CLES : consonnes géminées, japonais langue étrangère, perception, apprenants francophones.
KEYWORDS: geminate consonants, Japanese as a foreign language, perception, French-speaking learners.

1 Introduction

Les obstruantes géminées du japonais, ou *sokuon*, ont été abordées dans de nombreuses études, avec différentes perspectives. Les *sokuons* sont caractérisés par une durée (d'occlusion ou de bruit de friction) plus importante, environ deux fois plus que celle des équivalents simples (Kawahara, 2015). Ils se trouvent essentiellement en position intervocalique en milieu de mot, et de manière marginale en fin de mot (dans certaines interjections), mais jamais à l'initiale de mot, contrairement à certaines langues comme le tachelhit (Ridouane, 2007), le ryukyu d'Ōgami (Pellard, 2011), ou le dialecte d'Altamuro dans la province de Bari en Italie (Bertinetto, Loporcaro 2000). Il est également à noter que la durée de la voyelle précédant un *sokuon* est allongée en japonais (Fukui, 1978, entre autres), tandis que celle précédant une consonne géminée en italien, par exemple, est raccourcie, et que cette durée courte est utilisée par les auditeurs italophones pour distinguer une géminée d'une non-géminée (Esposito, Di Benedetto, 1999). Par ailleurs, des différences de mécanisme laryngien sont suggérées entre les *sokuons* et les obstruantes simples ; Fujimoto (2014) n'a observé aucune constriction laryngée apparente ou ni aucun coup de glotte, alors que des mesures de PGG (Photo-GlottoGraphe) et de kymographe montrent que l'ouverture de la glotte est restreinte au début du *sokuon* par rapport à une consonne simple, ce qui suggère une certaine tension des plis vocaux impliquée dans la production d'un *sokuon*.

Les *sokuons* correspondent ainsi grosso modo à des obstruantes longues sur le plan phonétique : les occlusives sourdes /p t k/ et la fricative /s/, et de façon marginale, la fricative /h/ essentiellement dans des interjections et onomatopées, et les obstruantes voisées /b d g z/ et le /r/ dans des emprunts récents. Sur le plan phonologique, les obstruantes géminées sont communément interprétées comme une séquence formée du phonème de gémination ou de *sokuon*, noté /Q/, suivi d'une obstruante (Vance, 2009, Labrune, 2012, entre autres).

Puisque qu'il s'agit d'une opposition distinctive, les apprenants du japonais langue étrangère / seconde ont besoin de la maîtriser afin de distinguer, entre autres, des paires minimales de formes verbales en *–te* (utilisées pour relier deux propositions, former l'impératif, …) comme /ˈkite/ (< /ˈkuru/ « venir ») vs. /ˈkiQte/ (< /ˈkiru/ « couper ») (notons qu'il existe également /kiQte/ « timbre », avec un accent lexical différent), ou /kaete/ (< /kaeru/ « changer, échanger ») vs. /ˈkaeQte/ (< /ˈkaeru/ « rentrer ») (malgré la différence d'accent lexical).

En ce qui concerne l'acquisition des géminées en japonais langue étrangère / seconde, de nombreuses études, portant notamment sur les apprenants anglophones, coréanophones et sinophones, suggèrent les difficultés avec lesquelles les apprenants parlant une langue sans distinction de la quantité consonantique produisent ou perçoivent les géminées : Sonu et al. (2012) sur la perception chez les apprenants coréanophones ; Tsukada et al. (2015) sur la perception chez les apprenants anglophones ; Hirata, Takiguchi (2015) sur la production chez les anglophones, pour ne citer que quelques études récentes. Ces derniers montrent que les *sokuons* produits par les apprenants anglophones du japonais ne sont pas suffisamment longs pour que les auditeurs natifs du japonais les perçoivent comme géminées. Les études empiriques sur l'acquisition des géminées par les apprenants francophones, cependant, semblent inexistantes. En nous fondant sur la littérature, nous pouvons prévoir une difficulté similaire, étant donné qu'une opposition de durée consonantique est absente en français, excepté les « fausses » géminées dues à la chute du /ə/ (ex. « pas de drap » /pa d dʁa/) ou une hypercorrection qui correspond à la graphie (ex. « collègue » [kɔllɛg]) : les deux catégories, obstruantes avec et sans *sokuon*, seraient perçues comme équivalentes d'une même catégorie phonémique en L1 (français), avec ou sans différence d'écarts

du prototype (PAM-L2 : Best, Tylor, 2007). L'objectif de cet article est de présenter une étude perceptive des obstruantes géminées du japonais chez des apprenants francophones.

2 Méthode

Afin d'étudier la perception de l'opposition de gémination en japonais, une tâche AXB a été assignée auprès d'apprenants francophones de japonais langue étrangères et d'auditeurs natifs du japonais.

Les stimuli utilisés dans cette expérience se composent de 30 non-mots dysyllabiques en japonais du type /b$V_iC_j(C_j)V_i$/, où la voyelle (V_i) est une des voyelles du japonais /i e a o u/, et la consonne simple (C_j) ou géminée (C_jC_j) est /p/, /k/, ou /s/ (Table 1). Notons que la syllabe /si/ se réalise phonétiquement [ɕi]. Tous les non-mots ont été prononcés par 2 locuteurs natifs du japonais de Tokyo (1 homme et 1 femme) avec deux accents lexicaux différents : HB (haut-bas) et BH. Les voyelles fermées /i/ et /u/ n'ont été dévoisées ni au milieu ni en fin de mot. Nous avons ainsi obtenu 30 paires minimales (3 consonnes x 5 voyelles x 2 accents lexicaux), avec ou sans gémination, qui ont été combinées en triplets. Si l'un(e) des deux locuteurs a été choisi(e) pour le deuxième stimulus de chaque triplet, l'autre a été retenu(e) pour les deux autres du même triplet. L'accent lexical (HB ou BH) était identique pour tous les stimuli du même triplet.

Non-mots **sans** *sokuon*			Non-mots **avec** *sokuon*		
bipi /bipi/	biki /biki/	bishi /bisi/	bippi /biQpi/	bikki /biQki/	bisshi /biQsi/
bepe /bepe/	beke /beke/	bese /bese/	beppe /beQpe/	bekke /beQke/	besse /beQse/
bapa /bapa/	baka /baka/	basa /basa/	bappa /baQpa/	bakka /baQka/	bassa /baQsa/
bopo /bopo/	boko /boko/	boso /boso/	boppo /boQpo/	bokko /boQko/	bosso /boQso/
bupu /bupu/	buku /buku/	busu /busu/	buppu /buQpu/	bukku /buQku/	bussu /buQsu/

TABLE 1 : Les 30 non-mots utilisés dans la tâche AXB : en translittération latine (style Hepburn) et transcription phonémique. Chaque mot a été prononcé avec deux accents lexicaux : HB et BH.

Trois groupes d'auditeurs ont participé à l'expérience : 1) FR1, composé de 10 étudiants de première année de licence d'études japonaises à l'INALCO (Institut National des Langues et Civilisations Orientales), âgés de 19 à 70 ans (26,9 ans en moyenne), qui avaient appris le japonais pendant une année universitaire, au moins (3,7 ans en moyenne) ; 2) FR2 : 9 étudiants de troisième année en études japonaises à l'INALCO, âgés de 21 à 25 ans (22,7 ans en moyenne), qui avaient étudié le japonais pendant 3 années universitaires, au moins (5,2 ans en moyenne) ; 3) JP : 6 auditeurs natifs du japonais, âgés de 21 à 38 ans (31,3 ans en moyenne).

Les auditeurs ont d'abord subi un entraînement afin de s'habituer à l'expérience. Ils devaient écouter les triplets avec un casque, décider si le deuxième mot de chaque triplet était identique au premier ou au dernier, et cliquer sur le bouton correspondant à leur réponse parmi ceux présentés sur un écran d'ordinateur. Une pause a été proposée après un bloc de 60 triplets. La présentation des stimuli et l'enregistrement des réponses ont été effectués avec Praat (Boersma, Weenink, 2015).

3 Résultats

Les taux d'erreur globaux sont représentés dans la figure 1 : 11,24% pour FR1, 10,55% pour FR2, 3,86% pour JP, respectivement en moyenne. La différence entre FR1 et JP ($\chi^2 = 45,69$; $p < 0,001$) et celle entre FR2 et JP ($\chi^2 = 38,35$; $p < 0,001$) s'avèrent significatives, tandis qu'il n'y a pas de différence significative entre les deux groupes d'apprenants FR1 et FR2 ($\chi^2 = 0,35$; $p = 0,55$).

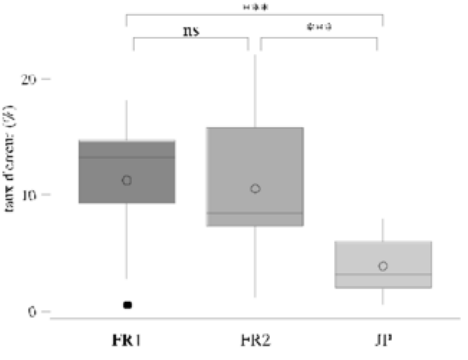


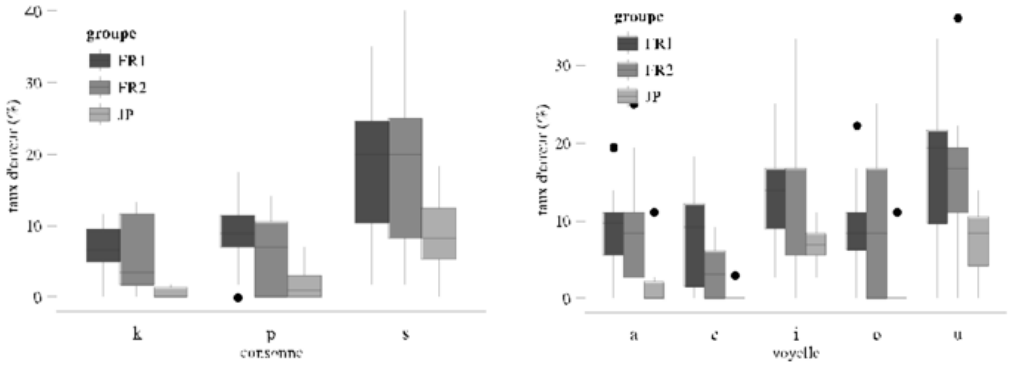
FIGURE 1: Taux d'erreur de discrimination AXB : FR1 (10 étudiants francophones de première année) ; FR2 (9 étudiants francophones de troisième année) ; JP (6 auditeurs natifs). 180 réponses par auditeur.

3.1 Résultats présentés par type de consonne

Les taux d'erreur calculés par type de consonne sont présentés dans la Figure 2. Les occlusives /p k/ ont induit moins d'erreurs (FR1 : 8,59% et 6,5% ; FR2 : 6,23% et 5,92% ; JP 2,04% et 0,55%, respectivement en moyenne pour /p/ et /k/) que la fricative /s/ (FR1 : 18,50% ; FR2 : 19,25% ; JP 8,89%, respectivement en moyenne). Il est à noter que le taux d'erreur s'élève à 8,89% même chez les auditeurs natifs (JP) pour /s/, tandis que peu d'erreurs ont été enregistrées pour les occlusives /p/ (2,04%) et /k/ (0,55%) chez ce même groupe d'auditeurs.

3.2 Résultats présentés par type de voyelle

La figure 3 montre les taux d'erreurs calculés par type de voyelle. Les voyelles fermées /i u/ ont fait produire plus d'erreur (FR1 : 13,33% et 16,39% ; FR2 : 14,19% et 16,05% ; JP 6,94% et 7,41%, respectivement en moyenne pour /i/ et /u/) que les autres voyelles (FR1 : 8,18%, 8,89% et 9,17% ; FR2 : 3,37%, 9,57% et 8,95% ; JP 0,5%, 2,31% et 1,85%, respectivement en moyenne pour /e/, /a/ et /o/). Notons également que les auditeurs natifs (JP) ont présenté des taux d'erreur non-négligeables pour /i/ et /u/ (6,94% et 7,41%, respectivement), alors que les erreurs étaient très peu nombreuses pour les autres voyelles.



FIGURES 2 - 3: Taux d'erreur de discrimination AXB calculés par type de consonne (/p k s/ : Fig. 2 à gauche), et par type de voyelle (/i e a o u/ : Fig. 3, à droite) : FR1 (10 étudiants francophones de première année) ; FR2 (9 étudiants francophones de troisième année) ; JP (6 auditeurs natifs). 60 (Fig. 2) et 36 (Fig. 3) réponses par consonne par auditeur.

3.3 Résultats présentés par type d'accent lexical

Les taux d'erreurs calculés par le type d'accent lexical (HB et BH) sont représentés dans la Figure 4 : ils sont plus élevés pour HB (FR1 : 14,83% ; FR2 : 14,18% ; JP 4,02%, respectivement en moyenne) que pour BH (FR1 : 7,78% ; FR2 : 7,04% ; JP 3,70%). Contrairement au cas des consonnes et de voyelles, les auditeurs natifs ont montré des taux d'erreurs similaires, quel que soit l'accent lexical.

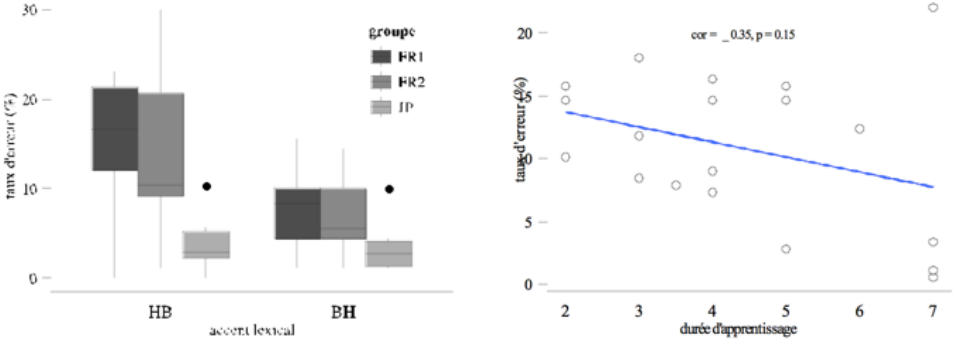


FIGURE 4 (à gauche) : Taux d'erreur de discrimination AXB calculés par type d'accent lexical (HB et BH) : FR1 (10 étudiants francophones de première année) ; FR2 (9 étudiants francophones de troisième année) ; JP (6 auditeurs natifs). 90 réponses par accent lexical par auditeur. FIGURE 5 (à droite) : Corrélation entre le taux d'erreur de discrimination AXB chez les 19 apprenants (FR1 et FR2 confondus) et leur durée d'apprentissage du japonais. 180 réponses par auditeur.

3.4 Corrélation entre le taux d'erreurs et la durée d'apprentissage

La figure 5 représente la corrélation entre le taux d'erreur chez les 19 apprenants (FR1 et FR2 confondus) et leur expérience d'apprentissage en nombre d'années. Le coefficient de Pearson étant de 0,35 ($p = 0,15$), la corrélation entre ces deux facteurs s'avère modérée.

4 Analyses statistiques

Jusqu'ici, nous avons vu les résultats de perception des *sokuons* en fonction des consonnes, des voyelles et du type d'accent lexical. Dans cette section, nous cherchons quel(s) facteur(s) agi(ssen)t sur cette perception. Pour cela, nous employons comme méthode statistique à déterminer le Modèle Mixte Linéaire Généralisé (MMLG / GLMM). GLMM est une méthode permettant de déterminer parmi plusieurs modèles celui qui explique le mieux les données observées. Il est possible de comparer des modèles incluant des nombres et des types de facteurs différents, qu'ils soient quantitatifs (durée d'apprentissage, âge de l'apprenant) ou qualitatifs (type de consonne, de voyelle et d'accent lexical).

Nous avons créé des modèles incluant différentes combinaisons des facteurs ci-dessus, mais en excluant les données des auditeurs japonais, car la durée de leur apprentissage de la langue n'est pas comparable avec celle des apprenants. Le GLMM retourne pour chaque modèle une valeur de AIC (*Akaike's Information Criteria*), une mesure de la qualité d'un modèle. Cette valeur permet ainsi de comparer les modèles et de déterminer celui qui prédit le mieux les données observées : le modèle avec le plus faible AIC est le plus vraisemblable. Les modèles varient en termes des facteurs quantitatifs inclus, mais les facteurs qualitatifs ont été conservés dans tous les cas. Nous n'avons pas inclus d'interactions entre les facteurs pour éviter la complexité d'interprétation des résultats. La Table 2 ci-dessous présente une comparaison des différents modèles testés.

	Modèle 1	**Modèle 2**	**Modèle 3**
Consonnes	+	+	+
Accent	+	+	+
Voyelles	+	+	+
Durée d'apprentissage	+	-	+
Âge	-	+	+
AIC	2046	2051	2048

TABLE 2 : les comparaison des modèles de GLMM

Le modèle 1 a le plus faible AIC, ce qui signifie que ce modèle, qui inclue comme facteur la durée d'apprentissage, est celui qui rend le mieux compte des données. Dans la section 3.4, nous avons examiné la corrélation entre le taux d'erreurs chez les 19 apprenants (FR1 et FR2 confondus) et leur durée d'apprentissage, et constaté qu'il n'y a pas de corrélation forte entre les deux. Néanmoins, d'après l'analyse de GLMM, nous pouvons dire que la durée d'apprentissage joue un rôle dans la perception des *sokuons*. Nous n'affirmons toutefois pas que la durée d'apprentissage

et l'âge sont les seuls facteurs expliquant la perception des *sokuons*. L'existence d'autres facteurs expliquant mieux cette perception reste à déterminer.

5 Discussion et conclusion

Les résultats de la tâche AXB montrent que la fricative /s/ pose plus de difficulté de distinction perceptive que les occlusives /p/ et /k/, et cette tendance est observée non seulement chez les apprenants mais aussi chez les auditeurs natifs. Selon Yanagisawa, Arai (2015), les transitions formantiques contribueraient à la distinction perceptive des géminées et des non-géminées chez les auditeurs japonophones natifs. Les transitions formantiques des fricatives moins abruptes (occlusion incomplète du conduit vocal) que celles des occlusives auraient ainsi rendu la tâche de discrimination plus difficile même chez les auditeurs natifs. Nos résultats seraient également dus au fait que la différence des durées de l'occlusion ou du bruit de friction entre les géminées et les simples est plus réduite pour la fricative /s/ que pour les occlusives (Kawahara, 2015).

Concernant les voyelles, les /i/ et /u/, qui ont induit plus d'erreurs chez les apprenants comme chez les natifs, sont connus pour leur durée plus courte que les voyelles non-fermées (Kaiki et al., 1992, entre autres), même si les contextes consonantiques sourds qui suscitent le dévoisement ont été évités dans la composition des non-mots. Avec le dévoisement, la tâche de discrimination serait encore plus difficile.

L'accent lexical est un autre facteur expliquant les résultats obtenus chez les apprenants. La hauteur tonale de chaque more (notons qu'un mot de deux syllabes avec un *sokuon* comporte trois mores) est communément décrite comme suit : HBB / HB et BBH / BH, avec et sans *sokuon*, respectivement (Saitô, 1997, entre autres). Même si la F0 est physiquement absente durant la consonne, le mouvement de F0 des mores adjacentes (notamment la more précédente) pourrait différencier la perception de la hauteur tonale, ce qui faciliterait la distinction dans le cas de BH (BB vs. BH sur les deux premières mores) que HB (HB vs. HB). Par ailleurs, la descente de F0 favoriserait la perception de *sokuon* (Kubozono et al., 2013), ce qui aurait induit plus de réponses pour le *sokuon* qu'il ne faut, d'où un taux plus élevé d'erreurs. L'explication de ces auteurs corrobore la tendance observée dans les tâches de dictée chez les apprenants francophones du japonais, qui ont tendance à insérer un *sokuon* à un mot prononcé sans *sokuon* plus fréquemment pour HB que pour BH (Tomoko Higashi, communication personnelle, 26 août 2015). L'influence de l'accent lexical est suggérée dans les résultats d'Ishizawa (2011), qui portent sur une tâche d'identification chez des apprenants anglophones.

Il est à noter ici que l'accent lexical semble influencer uniquement les résultats des apprenants, à la différence des consonnes et des voyelles, qui concernent aussi les auditeurs natifs (plus d'erreurs pour la fricative /s/ et les voyelles fermées /i/ et /u/). L'accent lexical étant une des difficultés majeures chez les apprenants francophones du japonais, il est possible que la différence de conscience de cette propriété ait pu induire les différences de perception : les natifs conscients auraient pu recourir à une certaine compensation perceptive, mais non les apprenants.

En ce qui concerne les participants et leur groupement, fondé, mis à part les auditeurs natifs, sur leur expérience d'études universitaires en japonais (première et troisième années de licence en études japonaises), les deux groupes d'apprenants n'ont pas montré une différence significative. Même s'il n'est pas certain que les capacités de discrimination des oppositions phonologiques soient en corrélation directe avec le niveau de compétences en compréhension orale ou dans d'autre domaines linguistiques, il faudrait tester d'autres critères de groupement susceptibles de

prédire les résultats, tels que le niveau obtenu en JLPT (*Japanese-Language Proficiency Test*) ou OPI (*Oral Proficiency Interview*), ou le résultat d'un autre test de compréhension orale, par exemple. Quant aux auditeurs natifs, un nombre plus élevé de participants, comparable à ceux des groupes d'apprenants, permettra une comparaison plus fiable.

Ultérieurement, il conviendra de comparer le cas du *sokuon* avec d'autres difficultés majeures chez les apprenants francophones (les voyelles longues, le /h/, entre autres). Par ailleurs, il sera important d'étudier les comportements des apprenants d'autres langues comme l'italien, par exemple, qui a une opposition phonémique de quantité consonantique (simple et géminée) similaire au système japonais, mais avec des réalisations phonétiques différentes, notamment avec une voyelle précédente raccourcie. Cela permettra de considérer des modèles de l'acquisition phonétique et phonologique des langues secondes, tel que le PAM-L2 (Best, Tylor, 2007), fondés sur l'influence du système phonologique des langues sources des apprenants.

Remerciements

Les auteurs remercient les participants au workshop « L'analyse des erreurs commises par les apprenants francophones du japonais langue étrangère » tenu à Bordeaux le 26 août 2015 pour leur commentaires sur une version antérieure de ce travail de recherche.

Références

BERTINETTO P. M., LOPORCARO M. (1999). Geminate distintive in posizione iniziale: uno studio percettivo sul dialetto di Altamura (Bari). *Annali della Scuola Normale Superiore di Pisa*, serie IV 4, 305-322.

BEST C. T., TYLOR, M. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. Bohn O.-S., Munro M. (éds.), *Language Experience in Second language Speech Learning. In honor of James Emil Flege*, Amsterdam: John Benjamins, 13-34.

BOERSMA P., WEENINK D. (2015). *Praat: doing phonetics by computer* [logiciel], Version 5.4.14 téléchargé en août 2015 depuis http://www.praat.org/.

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

FUJIMOTO M. (2014). Sokuon no kōtō chōsetsu no kōsokudo kamera to PGG ni yoru kentō : kēsu sutadī [Laryngeal Examination of Sokuon Using High-Speed Digital Video System and PGG: A Case Study]. *Journal of the Phonetic Society of Japan* 18(2), 44-53.

FUKUI S. (1978). Nihongo heisaon no enchō/tanshuku niyoru sokuon/hi-sokuon toshite no chōshu [Perception for the Japanese stop consonants with reduced and extended durations]. *The Bulletin of the Phonetic Society of Japan* 159, 9-12.

HIRATA Y., TAKIGUCHI I. (2015). Production of Japanese geminates by native English speakers: Durational accuracy and native speaker evaluation. Presented at *GemCon2015 ICPhS 2015 satellite workshop on "Geminate consonants across the world"*.

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.

ISHIZAWA T. (2011). Eigo o bogo to suru nihongo gakushūsha ni okeru nihongo sokuon no gochō: akusento to tango-nai no ichi ni chakumoku shite [An Error Analysis on the Perception of Japanese Geminate Consonants by Native English Learners of Japanese: From the perspective of accent and located position]. *Bulletin of the Graduate School of Education, Hiroshima University* 2-60, 173-181.

KAIKI N., TAKEDA K., SAGISAKA Y. (1992). Gengojōhō o riyō shita boin keizoku jikanchō no seigyo [Vowel duration control by linguistic information]. *IEICE (Institute of Electronics, Information and Communication Engineers) Transactions (Japanese Edition)* J-75A(3), 467-473.

KAWAHARA S. (2015). The phonetics of sokuon, or geminate obstruents. Kubozono H. (éd.), *Handbook of Japanese phonetics and phonology*, Berlin; Boston: De Gruyter Mouton, 43–78.

KUBOZONO H., TAKEYASU H., GIRIKO M. (2013). On the positional asymmetry of consonant gemination in Japanese loanwords. *Journal of East Asian Linguistics* 21(4), 339-371.

LABRUNE L. (2012). *The Phonology of Japanese*. Oxford: Oxford University Press.

PELLARD T. (2011). Ōgami (Miyako Ryukyuan). Shimoji M., Pellard T. (éds.), *An introduction to Ryukyuan languages*, Tokyo: Research Institute for Languages and Cultures of Asia and Africa (ILCAA), 113-166.

RIDOUANE R. (2007). Gemination in Tashlhiyt Berber: an acoustic and articulatory study. *Journal of the International Phonetic Association* 27(2), 119-142.

SAITÔ Y. (1997). *Nihongo onseigaku nyūmon [Introduction to Japanese Phonetics]*. Tokyo: Sanseidô.

SONU M., TAJIMA K., KATO H., SAGISAKA Y. (2012). Sokuon sōnyū handan ni chakumoku shita kankokugo bogo washa ni yoru nihongo sokuon no chikaku tokusē: kankokugo no nōon-ka to no kanrensei o chūshin ni [Perceptual characteristics of Japanese sokuon by Korean native listeners: Focusing on the relationship between sokuon insertion and Korean consonant tensification]. *Institute of Electronics, Information and Communication Engineers Technical Report* 111(471), 7-12.

TSUKADA K., COX F., HAJEK J., HIRATA Y. (2015). Perception of Italian and Japanese singleton/geminate consonants by listeners from different language backgrounds. Proceedings of *the 18th International Congress of Phonetic Sciences*.

VANCE T. (2009). *The Sounds of Japanese*. Cambridge: Cambridge University Press.

YANAGISAWA E., ARAI T. (2015). Forumanto sen'i to intensitī no gensui ga sokuon no chikaku ni ataeru eikyō [Influence of formant transitions and intensity attenuation on the perception of sokuon]. *Journal of the Acoustical Society of Japan* 71(10), 505-515.

La perception des séquences consonantiques non-natives par des locuteurs monolingues de mandarin

Qianwen Guan[1], Harim Kwon[1]
(1) CLLILAC-ARP, Université Paris Diderot, 75013 Paris
qianwen.guan@linguist.univ-paris-diderot.fr
harim.kwon@univ-paris-diderot.fr

RESUME

Cette étude examine le rôle de la structure phonotactique native et des facteurs phonétiques dans la perception des séquences consonantiques non-natives. Des locuteurs monolingues de mandarin ont été testés dans les deux expériences suivantes: dans la première expérience, les locuteurs ont du décider s'ils entendaient une voyelle entre deux consonnes en écoutant des séquences intervocaliques-CC (*a**kt**a*) et leurs contrôles CVC (*a**kat**a*). Les participants mandarins monolingues ont tendance à percevoir une voyelle entre deux consonnes dans les deux séquences CC et CVC. Mais le pourcentage de la voyelle perçue varie selon les différentes séquences. Dans la deuxième expérience, les mêmes participants ont écouté des séquences CC initiales et intervocaliques (***kt**apa, a**kt**a*) ainsi que CVC (***kat**apa*, *a**kat**a*) et les ont transcrites en Pinyin. Les stratégies observées dans la transcription: l'épenthèse, la métathèse, l'omission de C1 et celle de C2, montrent que les participants sont sensibles aux facteurs phonétiques. Les résultats des deux expériences suggèrent que la phonotactique native ainsi que des facteurs phonétiques affectent la perception des séquences non-natives.

ABSTRACT

Perception of non-native consonant sequences by Mandarin monolingual speakers

This study examines the role of native phonotactics and phonetic factors in the perception of non-native consonant sequences. Mandarin monolinguals were tested in two experiments. In Exp.1, the listeners were asked to decide if they heard a vowel between two consonants, after hearing non-native intervocalic CC sequences (*a**kt**a*) and CVC controls (*a**kat**a*). They reported hearing a vowel in both CC and CVC sequences, but to a varying extent for different consonant sequences. In Exp.2, the same listeners heard CC sequences in word-initial and intervocalic positions (***kt**apa, a**kt**a*) and their controls (***kat**apa, a**kat**a),* and transcribed them in Pinyin. The strategies observed in the transcription responses, including vowel epenthesis, metathesis, and consonant deletion, suggest the listeners are sensitive to phonetic details. Taken together, the results suggest both native phonotactics and phonetic factors influence perception of non-native consonant sequences.

MOTS-CLES: séquences consonantiques, perception, phonotactique, facteurs phonétiques
KEYWORDS: consonant sequences, perception, phonotactics, phonetic factors

1 Introduction

Des études précédentes montrent que lorsqu'on perçoit des mots non-natifs, on les adapte en

utilisant une stratégie systématique. Les locuteurs japonais, par exemple, ont tendance à percevoir une voyelle illusoire dans des séquences consonantiques (e.g. *ebzo -> ebuzo*) (Dupoux et al., 1999). Mais plusieurs études indiquent que les locuteurs ne perçoivent pas toutes les séquences non-natives avec le même pourcentage d'exactitude (Zsiga, 2003). Ces auteurs estiment que les facteurs phonétiques peuvent influencer la perception des séquences non-natives, par exemple, la durée du phonème et le délai de 'timing' articulatoire entre deux consonnes. Wilson et Davidson (2013) ont étudié la perception et la production des séquences initiales CC non-natives par des locuteurs natifs anglais et ils ont trouvé plusieurs stratégies d'adaptation. Le cas le plus fréquent est celui d'épenthèse vocalique entre les deux consonnes. Les auteurs attribuent cette adaptation à la sensibilité aux facteurs phonétiques acoustiques, comme la durée du relâchement de l'explosion des consonnes occlusives. Plus le relâchement est long, plus il favorise l'épenthèse vocalique. Zhao et Berent (2015) ont examiné la perception des séquences initiales par les locuteurs de mandarin et ont conclu que les facteurs phonétiques jouent un rôle important dans la perception des séquences non-natives. Mais les auteurs n'expliquent pas en détail ces facteurs phonétiques. De plus, les participants ne sont pas monolingues.

Notre étude a pour but de tester l'hypothèse du rôle de la structure phonotactique native et des facteurs phonétiques dans la perception des séquences consonantiques non-natives avec des locuteurs monolingues de mandarin. Comparé à l'anglais et au japonais, le mandarin a une structure syllabique plus simple: (C)(G)V(N). Il n'y a pas de séquences initiales-CC et la séquence 'nasale + occlusive' intervocalique est la seule séquence CC possible (inter-CC). Notre étude pose des nouvelles questions par rapport aux études précédentes: Sous quelles conditions les locuteurs perçoivent-ils les séquences CC avec une voyelle épenthétique ainsi que sans voyelle? Est-ce que leurs réponses dépendent du type de séquence consonantique? Pour répondre à ces questions, nous avons effectué une expérience d'identification des séquences consonantiques intervocaliques V*CC*V et une expérience de transcription pour comparer la voyelle en contexte intervocalique V*CC*V avec celle en contexte initial *CC*VCV. A notre connaissance, ces questions n'ont jamais été testées avec des locuteurs monolingues de mandarin.

2 Expérience 1. Identification

2.1 Méthode

Participants. 24 locuteurs monolingues de mandarin ont participé à cette expérience à Beijing. Le groupe est composé de 10 hommes et de 14 femmes, de 28 à 54 ans (la moyenne étant de 38 ans). Ils sont tous originaires du nord de la Chine et n'ont jamais vécu à l'étranger. Les dialectes de leurs régions n'ont pas de séquences consonantiques. Leur niveau d'éducation ne dépasse pas le lycée. Seulement 13 participants sur 24 ont des connaissances rudimentaires d'anglais. L'absence de familiarisation avec des séquences consonantiques a été testée et confirmée par la production de mots simples en anglais avec des séquences consonantiques: *stop, please, thanks,* etc. Lorsqu'ils prononcent ces mots, il y a toujours une voyelle épenthétique entre deux consonnes (FIGURE 1). Les participants n'ont jamais eu de troubles moteurs ni de parole, ni d'audition.

Matériels. Dans cette étude, nous avons utilisé 16 stimuli naturels, sans manipulation. Ce sont des non-mots, mais des mots possibles en russe, dont huit contiennent une séquence intervocalique V*CC*V (e.g. *áklu*) et huit autres sont des mots V*CVC*V avec une voyelle entre deux consonnes (e.g. *ákalu*). La première syllabe est toujours accentuée. La voyelle finale varie entre /u/ et /a/. Ce choix est fait par une locutrice native de russe, en fonction de ses intuitions de mots possibles en russe. Dans les mots retenus pour le contrôle, la voyelle /a/ entre les deux consonnes est une

voyelle non-accentuée réduite (Hamilton, 1980). C1 labiale est éliminée, parce que en mandarin, la voyelle épenthétique perçue dans les séquences C1-labial serait /u/, plutôt que la voyelle réduite /a/ dans le contrôle (Miao, 2005). La durée du relâchement de C1, s'il est présent, est mesurée selon l'oscillogramme et le spectrogramme du stimulus. La mesure inclut l'explosion de C1 et la friction qui suit (c.f. Wilson et Davidson, 2013).

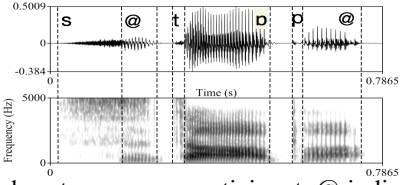


FIGURE 1: Spectrogramme de «stop» par un participant; @ indique la voyelle épenthétique /ə/.

V**CC**V (cible)	áklu	áknu	ákma	áksa	ákta	ákpa	átka	átpa
Relâchement C1	39	37	47	22	27	26	20	-
V**CVC**V (contrôle)	ákalu	ákanu	ákama	ákasa	ákata	ákapa	átaka	átapa
Relâchement C1	43	34	36	33	25	34	19	22

TABLEAU 1: Les stimuli et la durée (en ms) du relâchement C1. Dans /atpa/ le relâchement de [t] est faible et ne peut être mesuré.

Trois répétitions des stimuli ont été enregistrées par la locutrice native de russe dans une phrase porteuse. L'enregistrement a été réalisé avec une fréquence d'échantillonnage de 44.1 kHz dans une cabine isolée dans la salle d'expérimentation de l'Université Paris Diderot, à l'aide d'un ordinateur (Mac OS 10.10.1) avec le logiciel Praat version 5.4 (Boersma et Weenink, 2015), via un microphone AT2020. Nous avons ensuite sélectionné une répétition de chaque mot prononcé avec une intonation similaire, pour les utiliser dans l'Expérience 1.

Procédure. Les participants devaient indiquer s'ils entendaient une voyelle entre deux consonnes ou non. La présentation des stimuli et l'enregistrement des réponses ont été effectués sur un ordinateur PC avec un script Praat (ExperimentMFC). Les participants devaient appuyer sur la touche [z] du clavier du PC pour la réponse «oui» et sur la touche [?] pour la réponse «non». Nous avons utilisé 48 stimuli (8 séquences consonantiques * 2 structures (mot cible vs. contrôle) * 3 répétitions). Tous les stimuli ont été présentés de manière aléatoire et séparés par un ISI (Inter-Stimulus Interval) d'une seconde.

Prédictions. (1) Si la structure phonotactique joue un rôle dominant dans la perception des séquences non-natives, les participants pourront percevoir les séquences CC et les contrôles CVC sans faire de différence. (2) Si les facteurs phonétiques jouent un rôle dominant dans la perception des séquences CC, les participants pourront entendre une voyelle entre les deux consonnes plus souvent, c'est-à-dire, dès que le signal est plus similaire à un son vocalique. Plus précisément, les participants pourront percevoir plus de voyelles pour les stimuli dont les relâchements de C1 occlusive sont plus longs et ont plus d'énergie.

2.2 Résultats

Les participants ont perçu une voyelle entre deux consonnes dans CC stimuli dans 56% des cas et dans les contrôles CVC dans 71% des cas. La présence d'une voyelle perçue indique l'adaptation de la séquence CC par les locuteurs ainsi que la perception correcte pour les contrôles CVC.

Pour vérifier nos prédictions, les données ont été analysées à l'aide de modèles linéaires à effets mixtes dans lesquels les sujets et les stimuli figurent comme termes aléatoires. Les analyses ont été effectuées avec la fonction *glmer* du package *lme4* (Bates et al., 2014) du logiciel R (R Development Core Team, 2014). La réponse ('Voyelle' vs. 'Non-voyelle') est considérée comme variable dépendante, si les participants ont perçu une voyelle entre les séquences consonantiques. Le type de séquence consonantique (kl, km, kn, kp, ks, kt, tk, tp) et la structure (VCCV vs. VCVCV) sont incluses comme des facteurs fixes. Les interactions entre les deux facteurs ont aussi été testées. Un terme aléatoire 'Sujet' a permis de prendre en compte le cas d'une analyse à mesures répétées. Lorsqu'il y a eu un effet significatif ou une interaction, des analyses de Tukey's HSD post-hoc ont été réalisées en utilisant le package *lsmeans* (Lenth, 2015). Le seuil statistique a été fixé à $p < 0.05$.

Le type de séquence consonantique et le type de structure sont des facteurs significatifs des voyelles perçues par les participants. Le test post-hoc montre que le contrôle /akalu/ est plus souvent perçu avec une voyelle entre C1 et C2 que /aklu/ ($\beta = 2.50$, $p < 0.0001$). Quant aux autres séquences, la différence entre la séquence consonantique et le contrôle n'est pas significative.

Pour les séquences consonantiques sans voyelle entre les deux consonnes (VCCV), /kl/ est perçu avec une voyelle, significativement moins souvent que les autres séquences: /km, kn, ks, kp, kt, tk/ ($p < 0.05$), mais pas moins que /tp/ ($p = 0.33$). Aussi, /tp/ est moins souvent que /km, kn, ks, kt/ ($p < 0.05$), et /tk/ moins que /km, kt/ ($p < 0.01$). La différence entre /kp/ et /kt/ est marginalement significative ($p = 0.052$). Pour les contrôles avec une voyelle entre les deux consonnes (VCVCV), /tp/ est perçu avec une voyelle moins souvent que /km, kn, ks, kt, kp/ ($p < 0.01$), et /tk/ moins que /kn, kt/ ($p < 0.05$). Toutes les autres comparaisons ne sont pas significatives.

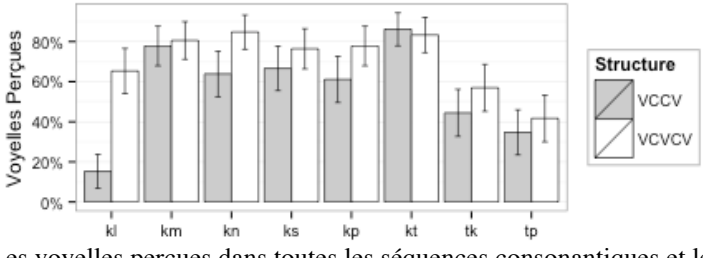


FIGURE 2: Les voyelles perçues dans toutes les séquences consonantiques et les contrôles. Les barres d'erreur indiquent l'intervalle de confiance à 95%.

2.3 Discussion

Dans l'Expérience 1, l'effet de la structure (CC vs. CVC) n'est pas significatif dans les séquences consonantiques, à l'exception de /kl/. Ceci confirme notre première prédiction sur le rôle de la structure phonotactique native dans la perception des séquences consonantiques non-natives. De plus, les participants ont perçu moins d'épenthèse avec C1-/t/ que C1-/k/. Wilson et Davidson (2013) ont montré que la durée du relâchement de C1 influence la perception d'une voyelle. Généralement, les relâchements de l'explosion des constrictions antérieures ont tendance à être plus faibles que ceux des constrictions postérieures (Stevens, 1998). De plus, /k/ a souvent une 'explosion double' dans nos stimuli. Ces deux conditions font que la durée du relâchement de /k/ est plus longue et l'énergie plus élevée dans l'explosion que pour /t/. Ce pattern s'applique à nos stimuli. Dans /atpa/, le relâchement de /t/ est masqué par la coarticulation dans le spectrogramme, et dans /atka/ et /akta/, la durée des relâchements de /t/ est plus courte que pour /k/ (FIGURE 3, voir

aussi TABLEAU 1). Une durée plus longue du relâchement pourrait favoriser la perception des voyelles épenthétiques par les locuteurs de mandarin.

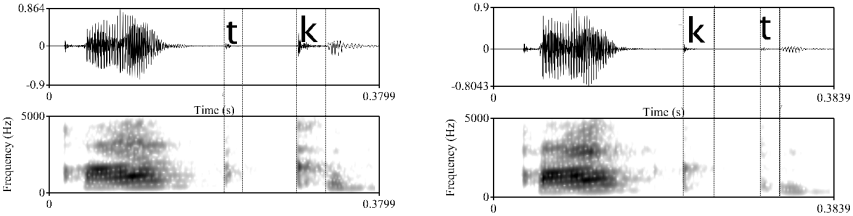


FIGURE 3: Spectrogramme de /atka/ (à gauche) et /akta/ (à droite) produits par la locutrice russe.

Ces résultats montrent que la structure phonotactique native et les facteurs phonétiques jouent un rôle non-négligeable dans la perception des séquences non-natives par les locuteurs de mandarin. Néanmoins, nous ne pouvons pas vérifier, dans cette expérience, si les participants ont utilisé d'autres stratégies que l'épenthèse. Par exemple, nous supposons que les participants n'ont pas perçu le /l/ sombre dans /aklu/. L'expérience de transcription présentée dans la section suivante expliquera par la suite les autres stratégies d'adaptation des séquences non-natives.

3 Expérience 2. Transcription

3.1 Méthode

Les participants de l'Expérience 1 ont ensuite transcrit en Pinyin les stimuli qu'ils ont entendus. Cette expérience contient 4 séquences consonantiques (*kl, kn, kt, tk*) dans deux positions (initiale *CC*VCV, intervocalique V*CC*V), et leurs contrôles équivalents avec une voyelle entre deux consonnes (initiale *CVC*VCV, intervocalique V*CVC*V) (le TABLEAU 3 montre tous les stimuli utilisés). Pour 16 non-mots, nous avons utilisé deux répétitions, sélectionnées des enregistrements de la même locutrice de russe de la première expérience. Le relâchement de C1 et C2 est mesuré pour chaque mot de la même façon que dans l'Expérience 1. La moyenne des durées pour chaque non-mot est présentée dans le TABLEAU 3.

Position \ Structure		Cibles (avec **CC**)				Contrôles (avec **CVC**)			
Initiale	Stimuli	**kl**ápa	**kn**ápa	**kt**ápa	**tk**ápa	**kal**ápa	**kan**ápa	**kat**ápa	**tak**ápa
	Relâchement C1/C2	64/ -	38/ -	39/15	26/28	38/ -	39/ -	29/16	17/36
Inter-vocalique	Stimuli	á**kl**u	á**kn**u	á**kt**a	á**tk**a	á**kal**u	á**kan**u	á**kat**a	á**tak**a
	Relâchement C1/C2	42/ -	43/ -	25/17	25/36	44/ -	36/ -	29/15	24/28

TABLEAU 3: Les stimuli et la durée moyenne (en ms) du relâchement de C1/C2

Prédictions. (1) Si la structure phonotactique native joue un rôle dans la perception des séquences non-natives, les participants pourront transcrire les séquences CC non-natives et leurs contrôles CVC sans différence. (2) Si les participants sont sensibles aux détails acoustiques en transcrivant les séquences non-natives en Pinyin, les transcriptions pourront montrer précisément les détails acoustiques des stimuli. Par exemple, plus le relâchement de C1 est saillant, plus le nombre de voyelles épenthétiques perçues entre deux consonnes sera grand (c.f. Wilson et Davidson, 2013). En revanche, si une consonne des séquences CC est moins bien perçue, les participants pourront la supprimer (omission de C1/C2).

3.2 Résultats

La stratégie d'adaptation la plus fréquente dans la perception des séquences consonantiques non-natives CC est la perception d'une voyelle épenthétique dans les deux positions: initiale-CC (70%) et inter-CC (80%). Comparés aux séquences CC, les pourcentages de perception d'une voyelle dans les contrôles CVC sont 99% en position initiale-CVC et 92% en position inter-CVC. Nous avons aussi observé d'autres stratégies d'adaptation: la métathèse (/atka/ -> /akəta/), l'omission de C1 (/atka/ -> /a_ka/) et l'omission de C2 (/atka/ -> /at_a/). Différentes stratégies peuvent être employées en même temps, comme pour /atka/ -> /akəta/. L'adaptation consiste ici en une épenthèse et une métathèse. Les pourcentages des quatre types d'adaptation dans la transcription des séquences initiales et intervocaliques CC et CVC ont été calculés en fonction de la composition des séquences consonantiques (FIGURE 4).

Les résultats ont été analysés à l'aide de quatre modèles à effets mixtes pour chaque type d'adaptation (Epenthèse, Métathèse, Omission de C1, Omission de C2). Pour chaque modèle, la variable dépendante était la présence/absence d'adaptation (Adaptation vs. Non-adaptation). Le type de séquence (kl, kn, kt, tk), la position (Initiale vs. Intervocalique) et la structure (CC vs. CVC) ont été inclus comme facteurs fixes, de même que les interactions entre eux.

Epenthèse. Les trois facteurs fixes ont tous des effets significatifs sur la perception d'une voyelle entre deux consonnes, de même que les interactions biunivoques entre la structure et la position, la structure et le type de séquence, la position et le type de séquence. Le test post-hoc montre que la structure CVC est perçue avec une voyelle plus souvent que la structure CC en position initiale-/kl/ ($\beta = 4.60$, $p < 0.0001$), initiale-/kn/ ($\beta = 5.06$, $p < 0.001$), initiale-/tk/ ($\beta = 3.00$, $p < 0.05$), inter-/kl/ ($\beta = 2.60$, $p < 0.0001$), et inter-/kn/ ($\beta = 3.06$, $p < 0.01$). De plus, les effets de la position sont significatifs pour les deux structures de /kl/: la position initiale-/kl/ est perçue avec une voyelle plus souvent que la position inter-/kl/ dans les mots cibles (/klapa/ vs. /aklu/, $p < 0.0001$), et dans les contrôles (/kalapa/ vs. /akalu/, $p < 0.0001$).

Omission de C1. L'interaction entre la structure et la position est significative. Le test post-hoc du HSD de Tukey montre que les effets de la position sont significatifs pour la structure CC ($\beta = 3.61$, $p < 0.05$), mais pas pour la structure CVC. Aucune des autres comparaisons n'est significative.

Omission de C2. Les effets des séquences consonantiques et de la position sont tous les deux significatifs, de même que les interactions biunivoques entre la structure et la position, la structure et le type de séquence, la position et le type de séquence. Le test post-hoc du HSD de Tukey montre que l'effet de la structure est seulement significatif pour inter-/kl/ (/aklu/ > /akalu/, $\beta = 3.22$, $p < 0.0001$). L'effet de la position est seulement significatif pour la séquence CC-/kl/ (/aklu/ > /klapa/, $\beta = 3.22$, $p < 0.0001$).

Concernant le type de séquence, /kl/ montre en général plus d'omission de C2 que les autres séquences, mais les effets sont un peu différents pour les positions et les structures différentes: en position initiale-C_1C_2, C_2 est moins perçue dans /klapa/ que dans /ktapa, tkapa/ ($p < 0.05$); en inter-C_1C_2, C_2 est moins perçue dans /aklu/ que dans /aknu, akta, atka/ ($p < 0.0001$), et dans /aknu/ que dans /atka/ ($p < 0.05$). Pour la structure CVC, les séquences /kl/ montrent plus d'omission de C2 que les autres séquences en position initiale et en position intervocalique ($p < 0.05$). Aucune des autres comparaisons n'est significative.

Métathèse. L'interaction entre la position et le type de séquence est significative: la séquence /tk/ a plus de métathèse que les autres séquences dans la position intervocalique uniquement ($p < 0.001$

pour /atka/, $p < 0.05$ pour /ataka/). De plus, l'interaction entre la structure et la position est significative: l'effet de la position est seulement significatif pour /atka/ vs. /tkapa/ ($\beta = 2.58$, $p < 0.0001$) et pour /ataka/ vs. /takapa/ ($\beta = 5.53$, $p < 0.05$).

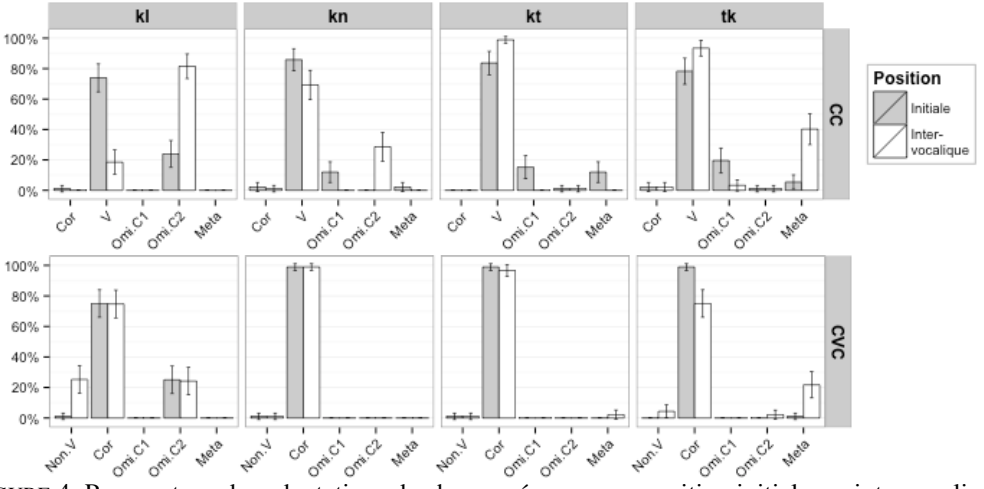


FIGURE 4: Pourcentage des adaptations de chaque séquence en position initiale vs. intervocalique, pour structures CC vs. CVC. «Cor» indique une perception correcte. Cela veut dire que dans la structure CC, les participants ont transcrit correctement une séquence CC et dans la structure CVC, les participants ont transcrit correctement une séquence CVC. «V» indique que les participants ont transcrit une voyelle entre deux consonnes dans une séquence CC. «Non.V» montre que les participants n'ont pas transcrit la voyelle dans une séquence CVC. «Omi.C1», «Omi.C2» et «Meta» signifient respectivement «Omission de C1», «Omission de C2» et «Métathèse». Les barres d'erreur indiquent l'intervalle de confiance à 95%.

3.3 Discussion

La comparaison des transcriptions des séquences consonantiques initiales et intervocaliques avec deux structures CC et CVC montre que les participants perçoivent très souvent une voyelle entre deux consonnes pour la plupart des séquences à quelques exceptions près (e.g. /aklu/). La FIG. 4 montre que la différence des stratégies d'épenthèse entre les deux structures CC et CVC n'est pas très grande. De plus, la stratégie de métathèse accompagne toujours la stratégie de l'épenthèse (/atka/ -> /akəta/). Ce résultat correspond aux résultats de l'Exp.1. Il confirme que la perception des séquences consonantiques non-natives est basée sur la structure phonotactique native.

Mais en analysant le détail des transcriptions, nous avons observé que les participants adaptent des stratégies différentes en fonction de facteurs phonétiques. En effet, l'absence d'épenthèse ne montre pas toujours que les participants ont entendu une séquence consonantique sans voyelle. De fait, il y a beaucoup de transcriptions avec omission de C1/C2, ce qui montre qu'ils n'ont pas entendu l'une des deux consonnes lorsque le signal acoustique est faible ou confus. Par exemple, il y a moins d'épenthèse et plus d'omission de C2 dans /aklu/ que dans /akalu/, parce que dans /aklu/ le /l/ sombre est plus souvent confondu avec la voyelle /u/ qui le suit. Comme on a vu dans la section 2.3, les formants du /l/ sombre ressemblent aux formants de la voyelle /u/ suivante au niveau acoustique (Sproat et Fujimura, 1993). D'ailleurs, l'intervocalique /aklu/ présente moins de voyelles épenthétiques et plus d'omission de C2 /l/ que l'initiale /klapa/. Cela veut dire que, le plus souvent, les participants n'ont pas perçu la liquide /l/ en position intervocalique, alors qu'ils la

perçoivent plus facilement en position initiale. Une explication pourrait être qu'en initiale de mot, les gestes consonantiques ont tendance à être moins co-produits, avec un délai plus long entre le geste consonantique de C1 et celui de C2 (Browman et Goldstein, 2001). Cela a été montré pour plusieurs langues, y compris le russe, la langue utilisée dans cette étude (Kochetov, 2006). De même, /n/ dans /aknu/ a été moins souvent perçu que les consonnes occlusives /t/ ou /k/. Acoustiquement, la nasale est plus proche d'une voyelle que les occlusives. Les occlusives se distinguent plus facilement de la voyelle suivante par leurs relâchements. Il semblerait que, pour les séquences consonantiques en position intervocalique, plus C2 est assimilée à la voyelle, moins elle est perçue, ce qui résulte en une omission de C2.

Le résultat montre aussi que l'initiale-C_1C_2VCV présente plus d'omission de C1 que l'intervocalique-VC_1C_2V. Ceci est dû au fait que dans les séquences intervocaliques, la transition formantique de C_1 est présente dans la voyelle précédente. Mais dans les séquences initiales, il y a moins d'information acoustique pour C_1 dans l'absence de la voyelle précédente. Les participants ont donc plus de possibilités de ne pas avoir perçu C_1 en position initiale qu'en position intervocalique.

Néanmoins, dans l'expérience de transcription, nous n'avons pas trouvé d'influence de la durée du relâchement de C1 sur la perception de l'épenthèse. Peu de locuteurs de mandarin ont transcrit correctement les séquences consonantiques, sans voyelle insérée. C'est peut-être parce que les séquences CC sont impossibles en Pinyin (sauf «nasale + occlusive» hétérosyllabique). Pourtant, une autre stratégie d'adaptation, la métathèse, pourrait être affectée par le relâchement de C1/C2. La métathèse de /tk/ en position intervocalique est présente dans les deux structures CC et CVC (/atka/, /ataka/ -> /ak*e*ta/). La présence de métathèse indique peut-être que les participants préfèrent /kt/ à /tk/, parce que /k/ est plus saillant que /t/. Nous confirmons que, dans nos stimuli, /k/ a une durée de relâchement plus longue que /t/ (TABLEAU 3). Son double relâchement lui donne souvent une énergie plus élevée que /t/. Selon Marin et al. (2010), quand /t/ et /k/ sont co-produits et donc ambigus, les locuteurs perçoivent la vélaire /k/ plus souvent que l'alvéolaire /t/. Ceci est dû au fait que la constriction du dos de la langue a en général plus d'influence sur l'acoustique (le relâchement) que celle du bout de la langue. De plus, Rice (1992) propose que la séquence /kt/ est préférée à /tk/ pour les séquences consonantiques hétérosyllabiques dans plusieurs langues, y compris le français (e.g. fa***ct***eur, mais *tk). La métathèse observée dans nos données pourrait avoir un lien avec cette préférence pour /kt/ plus que pour /tk/, parce qu'elle est présente significativement plus en position intervocalique qu'en position initiale.

4 Conclusion

Les résultats de notre étude confirment que la structure phonotactique native affecte la perception des séquences consonantiques non-natives par les locuteurs de mandarin. Parallèlement, ils soutiennent l'hypothèse que des facteurs phonétiques acoustiques jouent un rôle important dans la perception. Dans l'Expérience 1, des locuteurs monolingues de mandarin ont toujours perçu une voyelle dans les séquences intervocaliques CC (mots-cibles) et CVC (contrôles). De plus, le pourcentage de perception de la voyelle dans les séquences CC et CVC varie selon la durée du relâchement de C1. Dans l'Expérience 2, les participants ont principalement transcrit l'épenthèse comme stratégie d'adaptation. Par ailleurs, nous avons observé d'autres stratégies d'adaptation: l'omission des consonnes et la métathèse, qui sont influencées par les propriétés acoustiques de la consonne dans les séquences CC. Pour conclure, les résultats des deux expériences suggèrent que la structure phonotactique native ainsi que des facteurs phonétiques influencent la perception des séquences non-natives.

Remerciements

Cette étude a été réalisée avec le soutien financier du China Scholarship Council et de l'Institut des Etudes Doctorales de l'Université Paris Diderot (Bourse Doctorale de Mobilité Sortante).

Références

BATES D., MAECHLER M., BOLKER B., WALKER S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-7.

BERENT I., LENNERTZ T., JUN J., MORENO M. A., SMOLENSKY P. (2008). Language universals in human brains. *Proceedings of the National Academy of Sciences* 105(14), 5321-5325.

BOERSMA P., WEENINK D. (2015). *Praat: Doing Phonetics by Computer*. Version 5.4.22.

BROWMAN C., GOLDSTEIN L. (2001). Competing constraints on intergestural coordination and self-organization of phonological structures. *Bulletin de la Communication Parlée* 5, 25-34.

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, 1568-1578.

HAMILTON W. S. (1980). *Introduction to Russian Phonology and Word Structure*. Columbus: Slavica.

KOCHETOV A. (2006). Syllable position effects and gestural organization: Evidence from Russian. In L. Goldstein, D. H. Whalen & C. Best (eds.), *Laboratory Phonology 8*, 565-588.

LENTH R. V. (2015). *Least-Squares Means*. R package version 2.16.

MIAO R. Q. (2005). *Loanword Adaptation in Mandarin Chinese: Perceptual, Phonological and Sociolinguistic Factors*. PhD Dissertation, Stony Brook University.

R DEVELOPMENT CORE TEAM. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

RICE K. (1992). On deriving sonority: A structural account of sonority relationships. *Phonology* 9, 61-99.

SPROAT R., FUJIMURA O. (1993). Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of Phonetics* 21, 291-311.

STEVENS K. N. (1998). *Acoustic Phonetics*. Cambridge MA: The MIT Press.

WILSON C., DAVIDSON L. (2013). Bayesian analysis of non-native cluster production. In S. Kan, C. Moore-Cantwell & R. Staubs (eds.), *Proceedings of the Northeast Linguistics Society* 40, 265-278.

ZHAO X., BERENT I. (2015). Universal restrictions on syllable structure: Evidence from Mandarin Chinese. *Journal of Psycholinguistic Research* 44, 359-381.

ZSIGA E. C. (2003). Articulatory timing in a second language: Evidence from Russian and English. *Studies in Second Language Acquisition* 25, 399-432.

Perception et production de voyelles de l'anglais par des apprenants francophones : effet d'entraînements en perception et en production

Jennifer Krzonowski[1], Emmanuel Ferragne[2], François Pellegrino[1]
(1) Dynamique Du Langage – UMR 5596 – CNRS / Université Lyon 2
14 avenue Berthelot – 69007 Lyon
(2) CLILLAC-ARP EA 3967 / Université Paris Diderot
5 rue Thomas Mann – 75013 Paris
jennifer.krzonowski@cnrs.fr, emmanuel.ferragne@univ-paris-diderot.fr,
Francois.Pellegrino@univ-lyon2.fr

RESUME

Cette étude propose de tester l'effet de deux entraînements, en perception et en production, sur l'acquisition de voyelles de l'anglais britannique par des francophones. L'étude se focalise sur deux régions de l'espace acoustique pour lesquelles plusieurs catégories phonologiques existent en anglais alors qu'une seule existe en français. Trois groupes ont été constitués : l'un recevant un entraînement de type *High Variability Perceptual Training*, un second recevant un entraînement en production et le troisième constituait un groupe contrôle ne recevant pas d'entraînement. Les performances des participants ont été évaluées avant et après entraînement en perception et en production. Les résultats semblent montrer un effet de l'entraînement en perception sur les performances en perception et en production et un effet plus restreint de l'entraînement en production. Mais leur interprétation reste difficile du fait d'un effet test/re-test observé sur le groupe contrôle.

ABSTRACT

Perception and production of English vowels by French learners: effect of perception and production trainings

This study aims to evaluate the effect of two trainings, one in perception and one in production, on the acquisition of selected British English vowels by French learners. It focalizes on two regions of the acoustic space where there are several categories in English but only one in French. Three groups were formed: one receiving a *High Variability Perceptual Training*, the second one receiving a production training and the last one was a control group. Participants' accuracy in perception and production were assessed before and after training. The results seem to reveal a positive effect of perception training regarding perception and production performances, and a more restricted effect of the production training. But their interpretation remains difficult because of a test/re-test effect observed in the control group.

MOTS-CLES : Anglais langue seconde, contrastes phonologiques, voyelles de l'anglais, entraînement, perception, production

KEYWORDS: English as a second language, phonological contrasts, English vowels, training, perception, production

1 Introduction

L'acquisition d'une langue seconde (L2), en particulier à l'âge adulte, est toujours marquée par des difficultés en termes d'articulation et de perception des contrastes phonémiques non-natifs, c'est-à-dire des sons de la L2 qui n'existent pas ou ne sont pas phonologiquement distincts dans la langue maternelle (L1) des apprenants. Les apprenants de L2 sont donc souvent caractérisés par un accent étranger et une perception des sons de la L2 altérée (Strange et Shafer, 2008). Plusieurs facteurs participent à ces difficultés rencontrées dans l'acquisition d'une L2. Certains sont liés aux conditions de l'apprentissage, par exemple l'âge d'acquisition de la L2 ou les caractéristiques de l'exposition à la L2 (i.e., quantité et qualité). D'autres facteurs sont liés à des caractéristiques individuelles de l'apprenant comme la motivation ou les aptitudes dans l'apprentissage des langues (Piske, MacKay et Flege, 2001). Un autre facteur important est la relation entre les systèmes phonologiques de la L1 et de la L2.

En effet, de nombreuses études ont montré que la perception des sons de L2 est affectée par l'expérience linguistique et différents modèles explicatifs ont été développés, notamment le *Perceptual Assimilation Model* pour les apprenants de L2 (PAM-L2) (Best et Tyler, 2007) et le *Speech Learning Model* (SLM) (Flege, 1995). Selon ces deux modèles, les phonèmes de la L2 qui sont très différents des phonèmes existants en L1 seront assez bien perçus et prononcés ; en revanche, les phonèmes plus similaires à des catégories existant en L1 seront mal perçus et mal prononcés.

Considérant que les difficultés en perception de contrastes vocaliques non natifs constituent une part importante des problèmes rencontrés par les apprenants de L2, l'effet d'entraînements perceptifs des voyelles de l'anglais avec des locuteurs de diverses L1 a été étudié. Plusieurs études ont montré une amélioration significative des performances perceptives immédiatement après entraînement. Certaines études ont également montré une généralisation de l'apprentissage perceptif à de nouveaux locuteurs (e.g., Aliaga-Garcia, 2010; Nobre-Oliveira, 2007; Wang et Munro, 2004), à de nouveaux items (e.g., Aliaga-Garcia, 2010; Lacabex, Lecumberri et Cooke, 2009; Nobre-Oliveira, 2007) et à de nouveaux contextes (e.g., Aliaga-Garcia, 2010; Nobre-Oliveira, 2007), ainsi qu'une rétention pendant un mois (Nobre-Oliveira, 2007), deux mois (Rato, 2014) voire trois mois (Nishi et Kewley-Port, 2007; Wang et Munro, 2004) après la fin de l'entraînement. D'autres études ont comparé l'effet de différents types d'entraînements. Par exemple, Nobre-Oliveira (2007) a comparé l'utilisation de stimuli naturels et de stimuli de synthèse et n'a pas montré de différences significatives entre les deux types d'entraînements. Lacabex et al., (2009) et Aliaga-Garcia (2010) ont comparé des entraînements perceptifs et articulatoires (production) et montré que les deux types d'entraînements étaient efficaces pour améliorer les performances en perception.

L'influence d'autres variables a également été évaluée, notamment le niveau d'expertise en L2 (Iverson, Pinet et Evans, 2012) ou encore la taille des inventaires phonologiques en L1 et L2 (e.g., Iverson et Evans, 2009; Lengeris, 2009). Il a été montré que les entraînements étaient aussi efficaces quel que soit le niveau d'expertise en L2, et que l'apprentissage chez des locuteurs de L1 avec de plus grands inventaires phonologiques semble être facilité. Des apprenants dont la L1 présente un petit inventaire phonologique peuvent cependant très bien apprendre une L2 avec un plus grand inventaire. Les entraînements de type *High Variability Percetual Training* (HVPT), qui utilisent un grand nombre de stimuli produits par plusieurs locuteurs, ont montré leur efficacité sur les performances en perception (e.g., Aliaga-Garcia, 2010; Iverson et al., 2012; Wong, 2012). Quelques études se sont intéressées au transfert d'un entraînement en perception sur les performances en production, mais elles montrent toujours un effet supérieur de l'entraînement sur la perception (e.g., Lacabex et Lecumberri, 2010; Nobre-Oliveira, 2007).

À notre connaissance, seuls Iverson et al., (2012) ont testé l'effet d'entraînements perceptifs sur l'acquisition des voyelles de l'anglais par des apprenants francophones. Leur étude utilise un entraînement de type HVPT et leurs résultats, avec les participants non experts en anglais, montrent une faible amélioration des performances en perception et une amélioration en production limitée à certaines voyelles seulement. L'objectif de notre étude est de comparer l'effet de deux types d'entraînements, en perception et en production, sur les performances en perception et en production de voyelles de l'anglais par des apprenants francophones tardifs. L'entraînement proposé porte sur deux régions de l'espace vocalique qui ont la particularité de ne comporter qu'une seule catégorie en français mais deux (/ɪ - iː/) ou trois (/æ - ʌ - ɑː/) en anglais.

2 Méthode

2.1 Participants

48 participants volontaires de langue maternelle française, non bilingues, inscrits en première année d'anglais (LEA/LLCER) à l'Université Lyon 2 et Lyon 3 ont été divisés en 3 groupes. Deux groupes ont reçu un entraînement, soit en perception (groupe PE), soit en production (groupe PR) et un troisième groupe constituait le groupe contrôle (groupe C), qui n'a pas reçu d'entraînement mais dont les membres écoutaient des livres audio en anglais pour une durée équivalente aux entraînements. Chacun des groupes comportait 4 hommes et 12 femmes. Tous les participants signaient un formulaire de consentement au début de l'étude et étaient défrayés.

2.2 Matériel expérimental

Le matériel expérimental était constitué de paires minimales anglaises de type CVC portant sur deux régions de l'espace vocalique de l'anglais, la région /ɪ - iː/ et la région /æ - ʌ - ɑː/, pour laquelle les deux contrastes /æ - ʌ/ et /ʌ - ɑː/ ont été étudiés. Quatre paires ont servi dans les tâches des tests et dix autres ont servi dans les entraînements. Ces paires minimales ont été enregistrées par dix-huit locuteurs (neuf hommes et neuf femmes) de langue maternelle anglaise, originaires du Sud-Est de l'Angleterre. Les enregistrements ont eu lieu dans une salle isolée acoustiquement avec un micro USB Audio-Technica AT2020 avec le logiciel ROCme !. Le signal était converti au format numérique PCM mono avec un taux d'échantillonnage de 44,1 kHz et une résolution de 16 bits. Les signaux ont été analysés avec le logiciel Praat (Boersma et Weenink, 2015) afin d'en extraire des composantes temporelles et spectrales utilisées ensuite pour l'élaboration de l'entraînement en production. Les fichiers ont été segmentés manuellement pour ne contenir que le mot produit ; les trois premiers formants ainsi que la durée de la voyelle ont été mesurés par estimation semi-automatique : l'estimation formantique de Praat était superposée au spectrogramme, puis les seuils de détection étaient ajustés jusqu'à l'adéquation de l'estimation et du spectrogramme. Pour chaque formant, la valeur médiane sur la durée de la voyelle était ensuite calculée. Les stimuli ont été normalisés en amplitude. Les productions de six locuteurs (trois hommes et trois femmes) ont été choisies pour constituer le matériel des tests, celles des douze autres locuteurs (six hommes et six femmes) pour constituer le matériel des entraînements. Ainsi, 120 stimuli (5 voyelles × 4 paires minimales × 6 locuteurs) constituaient le matériel des tests, et le matériel des entraînements était composé de 600 stimuli (5 voyelles × 10 paires minimales × 12 locuteurs).

2.3 Procédure

2.3.1 Entraînements

Les entraînements comportaient cinq séances d'une heure maximum. Pour les deux types d'entraînement, en perception (groupe PE) et en production (groupe PR), les trois premières séances étaient focalisées sur une seule paire de voyelles (/ɪ - iː/, /æ - ʌ/, ou /ɑː - ʌ/) et les deux dernières séances portaient sur l'ensemble des voyelles. Les stimuli présentés dans les séances 1 à 3 contenaient toutes les paires minimales choisies pour constituer les stimuli d'entraînement. Afin de limiter la durée des séances 4 et 5, les stimuli ont été divisés en deux listes contenant chacune une moitié des paires minimales pour chaque contraste. Les entraînements ont été programmés avec l'interface *Demo Window* de Praat.

2.3.1.1 Entraînement en perception (Groupe PE)

L'entraînement en perception était un entraînement de type *High Variability Perceptual Training*. Il comportait des tâches d'identification (à deux choix forcés pour les séances 1 à 3 et à cinq choix forcés pour les séances 4 et 5) et des tâches de discrimination (*AX*, i.e., « identique ou différent » pour les séances 1 à 3, et *oddity*, i.e., détection d'intrus parmi trois items, pour les séances 4 et 5). Dans chacune des tâches, un feedback de type correct ou incorrect était donné après chaque item. En cas d'erreur, le participant réentendait l'item et devait choisir la bonne réponse.

2.3.1.2 Entraînement en production (Groupe PR)

L'entraînement en production était constitué d'une tâche de répétition de mots. À chaque item, le participant entendait un mot. Un indice visuel (i.e., symbole phonétique et mot indice) lui était donné quant à la voyelle contenue dans le mot. Puis il devait enregistrer sa production du mot. Les 3 premiers formants de la voyelle produite ainsi que sa durée étaient automatiquement mesurés. Pour cela, deux secondes de signal étaient enregistrées, sur lesquelles une détection de voisement était effectuée afin d'isoler la voyelle produite. Ensuite, un feedback visuel était présenté : la voyelle produite était représentée à l'écran dans le plan F1/F2 (point bleu). Au centre de l'écran un point rouge représentait la voyelle moyenne produite par les locuteurs natifs (enregistrés pour constituer le matériel) du même sexe que le participant. En bas de l'écran, la durée de la voyelle du participant était représentée (barre bleue) ainsi que la durée moyenne des voyelles correspondantes enregistrées par les locuteurs natifs de même sexe (barre rouge). Un feedback sur la qualité de la voyelle était également donné à partir d'une classification s'appuyant sur un modèle d'analyse discriminante appris sur les productions des locuteurs natifs de même sexe que le participant. Lorsque la voyelle produite était correctement classifiée et que la distance euclidienne à la cible dans le plan F1/F2 ne dépassait pas 200 Hz, la mention « CORRECT » était affichée à l'écran et le mot suivant était présenté. Dans le cas contraire, « Please Try again » était affiché si le participant n'avait pas dépassé deux essais et le même mot était présenté à nouveau. Après le troisième essai, « INCORRECT, go to the next word » apparaissait et le participant passait au mot suivant.

2.3.2 Tests (un jour avant et un jour après l'entraînement, T1 et T2)

2.3.2.1 Identification

Les participants réalisaient une tâche d'identification à cinq choix forcés. Les mots étaient présentés en ordre aléatoire au moyen d'un casque audio. La tâche comportait 120 items (5 voyelles × 4 contextes CVC × 6 locuteurs (3 hommes et 3 femmes)). Les participants devaient indiquer quelle voyelle parmi 5 (/ɪ/, /iː/, /æ/, /ʌ/, ou /ɑː/) était prononcée en cliquant sur un bouton à l'écran. Les

boutons à l'écran contenaient le symbole phonétique de la voyelle ainsi qu'un mot indice fréquent contenant la voyelle (*pig, feet, cat, cup et card*). Les participants ne recevaient aucun feedback et ne pouvaient pas réécouter le mot.

2.3.2.2 Discrimination

Les participants réalisaient ensuite une tâche de discrimination de type *Oddity*. À chaque item les participants entendaient, au moyen d'un casque audio, une séquence de trois mots CVC anglais prononcés par 3 locuteurs différents de même sexe à un intervalle inter stimuli de 500 ms ; deux mots étaient identiques et un différent. Le participant devait décider quel mot était l'intrus en cliquant sur l'un des trois boutons (labellisés 1, 2 ou 3) affichés à l'écran. Pour chaque paire de voyelles (/ɪ - i:/, /æ - ʌ/, ou /ɑ: - ʌ/), 4 paires minimales (e.g., *bid – bead, bit – beat, hid – heed, hit – heat*) étaient présentées 6 fois chacune (e.g, la moitié avec *bit*, l'autre moitié avec *beat* comme intrus en première, deuxième ou troisième position). La tâche comportait 144 items (3 contrastes × 4 paires minimales × 6 combinaisons × 2 sexe de locuteur) présentés en ordre aléatoire.

2.3.2.3 Production

Les participants produisaient isolément des mots anglais de type /bVd/ contenant les voyelles /æ/ (*bad*), /ʌ/ (*bud*), /ɑ:/ (*bard*), /i:/ (*bead*), et /ɪ/ (*bid*). Les mots étaient présentés avec le logiciel ROCme !, 3 fois chacun, dans un ordre aléatoire. Après cinq items d'essai constitués de mots fréquents contenant ces voyelles (*cat, cut, card, feet* et *pig*), les mots à produire étaient présentés visuellement à l'écran accompagnés du symbole phonétique de la voyelle contenue dans le mot et du mot fréquent pour la voyelle présenté en essai.

3 Résultats

3.1 Identification

Une première analyse de la variance sur les proportions d'identification correcte à T1 après transformation arc-sinus ne montre pas d'effet du facteur Groupe indiquant que les trois groupes (PE, PR et C) ne diffèrent pas significativement avant entraînement. Une analyse de la variance sur les différences des proportions d'identification correcte après transformation arc-sinus entre T1 et T2 montre un effet significatif du facteur Groupe $F(2, 45) = 7.15$, $p = .002$. Il n'y a pas d'effet du facteur Voyelle, ni d'interaction Groupe × Voyelle. Des comparaisons post-hoc entre les groupes montrent une différence significative entre les groupes C et PE ($p = 0.002$), une différence marginale entre les groupes C et PR ($p = 0.06$) mais pas de différence significative entre les deux groupes expérimentaux PE et PR. Bien que l'interaction Groupe × Voyelle ne soit pas significative, des *t* de Student pour mesures appariées réalisés par groupe pour chaque voyelle testée entre les scores à T1 et à T2 montrent que les participants du groupe PE augmentent significativement leurs performances entre T1 et T2 pour toutes les voyelles, que les participants du groupe PR améliorent significativement leurs performances entre T1 et T2 pour toutes les voyelles sauf le /ɑ:/. Enfin, le groupe C montre une amélioration significative des performances en discrimination pour les voyelles /ʌ/ et /ɪ/. Ces résultats sont présentés dans la FIGURE 1.

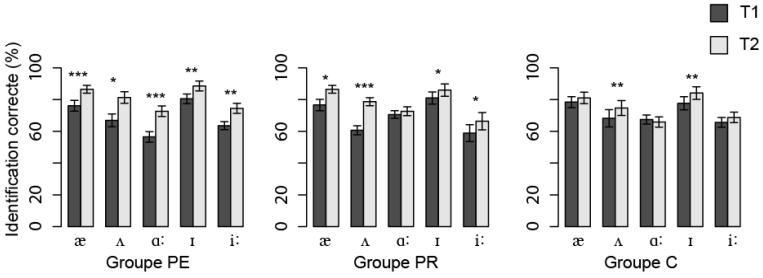


FIGURE 1 : Pourcentages moyens d'identification correcte à T1 (gris foncé) et à T2 (gris clair) selon le groupe pour chaque voyelle. (Les astérisques indiquent le niveau de significativité des différences : * p < .05 ; ** p < .01 ; *** p < .001)

3.2 Discrimination

Une première analyse de la variance sur les proportions de discrimination correcte à T1 après transformation arc-sinus ne montre pas d'effet du facteur Groupe indiquant que les trois groupes (PE, PR et C) ne diffèrent pas significativement avant entraînement. Une analyse de la variance sur les différences entre les proportions de discrimination correcte après transformation arc-sinus entre T1 et T2 montre un effet significatif du facteur Groupe $F(2, 45) = 5.00$, $p = .01$. Il n'y a pas d'effet du facteur Contraste, ni d'interaction Groupe × Contraste. Des comparaisons post-hoc entre les groupes montrent une différence significative entre les groupe PE et C ($p = .008$), mais pas de différence entre les groupes PR et C, ni entre les deux groupes expérimentaux PE et PR. Bien que l'interaction Groupe × Contraste ne soit pas significative, des *t* de Student pour mesures appariées réalisés par groupe pour chaque contraste montrent que les deux groupes expérimentaux, i.e., PE et PR, améliorent significativement leurs performances en discrimination pour les trois paires de voyelles testées, alors que le groupe C améliore ses performances seulement pour les paires /æ - ʌ/ et /ɑ: - ʌ/. Ces résultats sont présentés dans la FIGURE *2*.

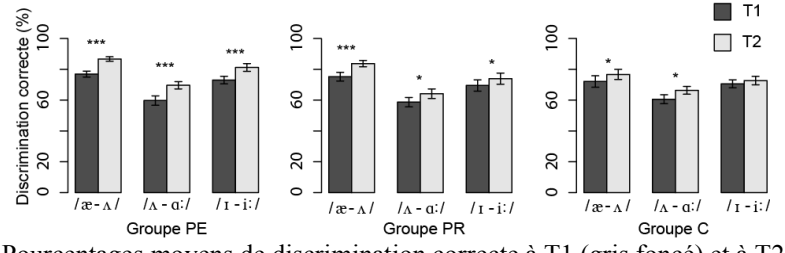


FIGURE 2 : Pourcentages moyens de discrimination correcte à T1 (gris foncé) et à T2 (gris clair) selon le groupe et pour chaque contraste. (Les astérisques indiquent le niveau de significativité des différences : * p < .05 ; ** p < .01 ; *** p < .001)

3.3 Production

Les deux premiers formants et la durée des voyelles produites ont été mesurés de la même manière que pour le matériel expérimental. Pour chaque production, à T1 et T2, la médiane, calculée sur la durée de la voyelle, de chacun des deux premiers formants a été extraite en Hertz, transformée en Bark (Traunmüller, 1990) et centrée-réduite indépendamment pour chaque formant et pour chaque participant. Ces valeurs ont ensuite été moyennées par type de voyelle (3 occurrences chacun) pour chaque participant à T1 et T2. À partir de ces valeurs de formants (F1 et F2), une matrice des distances

entre les cinq voyelles testées a été calculée pour chaque apprenant et comparée aux matrices de distances des voyelles des locuteurs natifs par le biais de corrélations de matrices. Considérant qu'une matrice des distances des voyelles donne une idée de la structure de l'espace des voyelles, une forte corrélation entre les matrices des apprenants et des locuteurs natifs devrait indiquer une grande proximité des systèmes vocaliques (Ferragne et Pellegrino, 2007). Nous avons donc mesuré l'effet des entraînements par le biais de la variation des coefficients de corrélation *r* des matrices des apprenants et des locuteurs natifs. Une première analyse de la variance sur le coefficient de corrélation de matrices à T1 montre un effet principal du Groupe, $F(2, 45) = 9.91$, $p < .001$. Des comparaisons post-hoc montrent qu'à T1, les groupes PR et C ont des performances significativement plus élevées que le groupe PE ($p < .001$) et que le groupe C a également des performances significativement plus élevées que le groupe PR ($p = .001$). Une analyse de la variance sur cette même mesure incluant le facteur Test montre toujours un effet significatif du facteur Groupe $F(2, 45) = 10.4$, $p < .001$. Cependant, nous n'observons pas d'effet significatif du Test, ni d'interaction Groupe × Test. Néanmoins, des *t* de Student pour mesures appariées réalisés sur chacun des groupes montrent que les deux groupes expérimentaux présentent un indice de corrélation plus élevé à T2 qu'à T1, ce qui suggèrerait que la structure de l'espace des voyelles testées se rapproche de celle de locuteurs natifs après entraînement (FIGURE 3). Ces résultats restent difficiles à interpréter puisque le groupe C présentait, déjà à T1, des scores élevés.

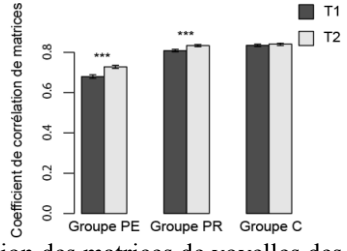


FIGURE 3 : Coefficient de corrélation des matrices de voyelles des apprenants et des locuteurs natifs, à T1 (gris foncé) et à T2 (gris clair) pour chaque groupe. (Les astérisques indiquent le niveau de significativité dans différences : * p < .05 ; ** p < .01 ; *** p < .001)

4 Discussion

L'objectif de cette étude était de déterminer l'effet d'entraînements en perception et en production sur l'acquisition de certaines voyelles de l'anglais par des apprenants francophones, les voyelles /ɪ/, /i:/, /æ/, /ʌ/, et /ɑ:/. L'acquisition de ces voyelles est difficile pour les francophones, car elle implique une modification de leur système phonologique de manière à créer de nouvelles catégories dans des régions de l'espace vocalique où des catégories existent déjà en français. Selon le PAM et le SLM, ils vont avoir tendance à assimiler ces voyelles à celles existant dans leur L1, leur rendant la discrimination de ces contrastes difficile (Best et Tyler, 2007; Flege, 1995) et entraînant des difficultés dans la production des phonèmes de L2.

Deux types d'entraînements ont été proposés, un en perception (groupe PE), un autre en production (groupe PR). Les performances des participants, en perception et en production, ont été évaluées avant et après entraînement et comparées à celles d'un groupe contrôle (groupe C) ne recevant pas d'entraînement. Ainsi, l'idée était d'observer les effets des entraînements sur les performances des participants dans la modalité entraînée (perception ou production), mais également d'observer un transfert de ces effets vers l'autre modalité. Les résultats montrent que le groupe PE semble avoir globalement bénéficié de l'entraînement en ce qui concerne les habiletés en perception. Concernant

les performances de ce groupe en production, l'amélioration observée est plus difficile à interpréter car elle ne diffère pas significativement de celle observée dans le groupe C. L'effet de l'entraînement reçu par le groupe PR est difficile à interpréter également, puisque ce groupe se distingue bien du groupe C dans la tâche d'identification, mais pas dans la tâche de discrimination, ni dans les performances en production. De plus, le groupe C montre également une amélioration de ses performances en perception (en identification et en discrimination), qui peut être interprétée comme un effet test/re-test. On n'observe pas de tel effet pour ce groupe sur les performances en production, mais il faut noter que les performances des participants de ce groupe sont déjà élevées en pré-test. Ainsi, l'interprétation des effets observés sur les groupes expérimentaux reste difficile à interpréter.

Ces résultats et leur interprétation soulèvent une question abordée par Halliday (2014) concernant l'évaluation des effets d'entraînements. L'auteur compare deux études présentant des protocoles d'entraînements similaires mais avec des conclusions différentes quant à l'efficacité des entraînements. En effet, dans l'une des deux études, le groupe contrôle utilisé montre une amélioration de ses performances plus importante que dans l'autre étude, autrement dit, un effet test/re-test plus important, ce qui vient réduire l'effet de l'entraînement observé pour cette étude.

On peut questionner également les mesures choisies pour évaluer l'amélioration des performances, tant en perception, qu'en production. Pour les performances en perception, les scores utilisés ici sont fréquemment utilisés dans la littérature, néanmoins, il serait nécessaire de comparer les scores obtenus par les différents groupes à ceux de locuteurs natifs, afin de repérer des effets plafonds qui peuvent être à l'origine de l'absence d'effet des entraînements pour certains apprenants. Il en est de même pour les performances en production ; dans notre cas, il serait intéressant de recueillir l'évaluation des productions des apprenants avant et après entraînement par des locuteurs natifs. La mesure que nous avons choisi d'utiliser ici (i.e., coefficient de corrélation *r* entre matrices de distances entre voyelles des apprenants et de locuteurs natifs) est censée refléter un degré de proximité du système des apprenants et de celui des natifs, et ainsi donner une idée de la justesse de réalisation des voyelles par les apprenants. Mais il n'est pas certain qu'elle reflète réellement la justesse de réalisation perçue par un locuteur natif.

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

ALIAGA-GARCIA, C. (2010). Measuring perceptual cue weighting after training: A comparison of auditory vs. articulatory training methods. Acte de *New Sounds 2010*, Poznan, Poland.

BEST, C. T. et 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*, 13-34.

BOERSMA, P. ET WEENINK, D. (2015) Praat: doing phonetics by computer. (Version 5.4.22). Repéré à www.praat.org

FERRAGNE, E. et PELLEGRINO, F. (2007). Automatic dialect identification: A study of British English *Speaker classification II* (p. 243-257): Springer.

FLEGE, J. E. (1995). Second Language Speech Learning Findings and Problems. Dans W. Strange (dir.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. Timonium, MD: York Press.

HALLIDAY, L. F. (2014). A tale of two studies on auditory training in children: A response to the claim that 'discrimination training of phonemic contrasts enhances phonological processing in mainstream school children, by Moore, Rosenberg and Coleman (2005). *Dyslexia, 20*(2), 101-118.

IVERSON, P. et EVANS, B. G. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *Journal of the Acoustical Society of America, 126*, 866-877.

IVERSON, P., PINET, M. et EVANS, B. G. (2012). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics, 33*(01), 145-160.

LACABEX, E. G. et LECUMBERRI, M. (2010). Investigating training effects in the production of English weak forms by Spanish learners Acte de *New Sounds 2010*.

LACABEX, E. G., LECUMBERRI, M. L. G. et Cooke, M. (2009). Training and generalization effects of English vowel reduction for Spanish listeners. Dans M. A. Watkins, A. S. Rauber & B. O. Baptista (dir.), *Recent Research in Second Language Phonetics/Phonology: Perception and Production* (p. 32-42). Newcastle upon Tyne: Cambridg Scholars Publishing.

LENGERIS, A. (2009). *Individual differences in second-language vowel learning.* (University College London, London).

NISHI, K. et KEWLEY-PORT, D. (2007). Training japanese listerners to perceive American English vowels: Influence of training sets. *Journal of Speech, Language, and Hearing Research, 50*, 1496-1509.

NOBRE-OLIVEIRA, D. (2007). *The effect of perceptual training on the learning of english vowels by brazilian portuguese speakers.* (Universidade federal de santa catarina, Florianopolis).

PISKE, T., MACKAY, I. R. et FLEGE, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics, 29*(2), 191-215.

RATO, A. (2014). Effects of Perceptual Training on the Identification of English Vowels by Native Speakers of European Portuguese. Acte de *Concordia Working Papers in Applied Linguistics*.

STRANGE, W. et SHAFER, V. L. (2008). Speech perception in second language learners: The re-education of selective perception. Dans J. E. Hansen Edward & M. L. Zampini (dir.), *Phonology and second language acquisition* (Vol. 36, p. 153-192). Philadelphia: John Benjamins.

TRAUNMÜLLER, H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America, 88*(1), 97-100.

WANG, X. et MUNRO, M. J. (2004). Computer-based training for learning English vowel contrasts. *System, 32*, 539-552.

WONG, J. W. S. (2012). Training the perception and production of English /e/ and /ae/ of Cantonese ESL learners: A comparison of low vs. high variabilité phonetic training. Acte de *14th Australian International Conference on Speech Science and Technology*.

Perception native des voyelles catalanes produites par des locutrices multilingues

Cynthia Magnen[1] Josefina Carrera-Sabaté[2] Pascal Gaillard[3]
(1) MSHS-T USR3414, Université de Toulouse et CNRS, Toulouse, France
(2) GRAN VIA DE LES CORTS CATALANES, 585, 08007, Universitat de Barcelona, Barcelona, Espagne
(3) CLLE UMR5263, Université de Toulouse et CNRS, Toulouse, France
cynthia.magnen@univ-tlse2.fr, jcarrera@ub.edu,
pascal.gaillard@univ-tlse2.fr

RESUME

Cette étude porte sur les voyelles catalanes produites par des adolescentes multilingues en Catalan-Castillan ayant pour langue maternelle soit le Catalan, soit le Roumain, soit l'Arabe du Maghreb. Nous proposons à vingt-et-un auditeurs catalanophones natifs un Test de Catégorisation Libre des voyelles produites dans ce contexte multilingue. Ce faisant, nous testons le modèle *Automatic Selective Perception* (ASP - Strange, 2011) qui stipule qu'en fonction de la variabilité des stimuli et de la tâche proposée, les auditeurs réalisent un traitement des stimuli selon un mode phonétique ou phonologique. Les résultats indiquent que le traitement des stimuli est double : les voyelles moyennes sont traitées selon un mode phonétique, tandis que les voyelles extrêmes sont traitées selon un mode phonologique. L'assimilation de voyelles d'une catégorie vocalique à une autre informe sur la qualité des réalisations non natives et témoigne de l'influence de la L1.

ABSTRACT

Native perception of Catalan vowels uttered by female multilingual speakers.
This study deals with Catalan vowels produced by adolescent multilingual speakers of Catalan-Spanish who have either Catalan, Romanian or Maghrebi Arabic as their native language. We gave 21 native Catalan speakers a free sorting task based on the vowels produced in this multilingual context. This was to test the Automatic Selective Perception model (Strange, 2011), which says that, depending on the variability of the stimuli in the task, listeners will treat the stimuli either in a phonetic mode or in a phonological mode. The results obtained indicate that dual treatment was given to the stimuli: mid vowels were treated in a phonetic mode, while point vowels were treated in a phonological mode. The assimilation of vowels from one category into another provides information on the quality of non-native utterances and bears witness to the influence of speakers' L1.

MOTS-CLES : Perception, Catalan, catégorisation, contraste non natif, multilinguisme.
KEYWORDS: Perception, Catalan, categorization, non-native contrast, multilingualism.

1 Introduction

Se présentant comme les deux langues les plus parlées en Catalogne, le catalan et l'espagnol partagent le statut de langue officielle de la région. Pourtant la Catalogne n'échappe pas au phénomène croissant d'arrivées massives de populations migrantes. La ville de Lleida, située dans la partie ouest de la région catalane, fait partie des villes où l'augmentation est la plus élevée. Les langues étrangères parlées à Lleida peuvent être divisées en quatre groupes : les langues arabes et

berbères, les langues latino-américaines, les langues subsahariennes et les langues d'Europe de l'Est (*cf.* Lorés *et coll.*, 2010). La situation linguistique de la Catalogne et en particulier de la ville de Lleida offre un cadre riche et parfaitement adapté à l'étude des productions non natives. Ainsi, des travaux menés par Carrera-Sabaté (2013) ont porté sur les productions de voyelles catalanes par des locuteurs adolescents multilingues d'origine roumaine et arabe et ayant migré en Catalogne pendant l'enfance. Dans la section suivante, nous présentons les résultats de mesures formantiques des réalisations vocaliques catalanes produites par ces groupes de locuteurs.

1.1 Productions non natives de voyelles catalanes par des locuteurs multilingues natifs de Roumanie et du Maghreb

La figure 1 synthétise les résultats d'études comparatives portant sur la production de 5068 exemplaires de voyelles catalanes [i e ɛ a o ɔ u]. Ces voyelles ont été produites par 39 locuteurs (20 filles et 19 garçons) au sein de logatomes (de forme CVC où C1=C2 et V est accentuée). Les locuteurs étaient des adolescents âgés entre 12 et 17 ans, bilingues Catalan-Castillan. Tous vivaient et étaient scolarisés dans la ville de Lleida. Trois sous-groupes se distinguaient parmi ces locuteurs en fonction de leur langue d'origine : ceux-ci avaient soit le Roumain (RM), soit l'Arabe (ARB), soit le Catalan (CAT) comme langue maternelle (L1).

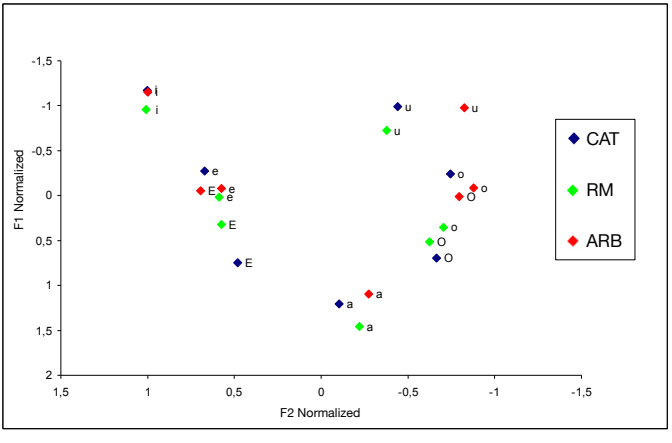


FIGURE 1 : Voyelles accentuées du Catalan produites par des adolescents dont la L1 est soit le Catalan (CAT), le Roumain (RM) ou l'Arabe (ARB) (Carrera-Sabaté, 2013).

Parmi les différences les plus importantes entre les réalisations des différents groupes interrogés, Carrera-Sabaté (2013) a mis en évidence des difficultés à produire le contraste entre les voyelles moyennes [ɛ-e] et [ɔ-o] du Catalan pour les catalanophones non natifs (*cf.* Figure 1). En effet, l'écart entre les valeurs de F1 des voyelles mi-fermées et mi-ouvertes est peu important pour les deux groupes de locuteurs catalanophones non natifs (ARB et RM) suggérant que le contraste n'est pas ou peu produit. Toutefois, des différences dans les réalisations des voyelles moyennes par les deux groupes de locuteurs non natifs ont été relevées. Concernant les réalisations des locuteurs arabophones, les valeurs de F1 des voyelles [ɛ-e] et [ɔ-o] sont très proches des voyelles fermées [e] et [o] du Catalan. La tendance inverse est observable dans la réalisation du contraste [ɔ-o] par les locuteurs roumanophones puisque les valeurs de F1 des deux voyelles de ce contraste s'approchent de la voyelle ouverte [ɔ] produite par les catalanophones. Par ailleurs, on observe que les productions

roumanophones et catalanophones de la voyelle [e] sont très proches tandis que les valeurs de F1 de la voyelle [ɛ] se trouvent entre les voyelles [ɛ] et [e] produites par les catalanophones.

Les tendances divergentes observées dans la réalisation des voyelles moyennes par les deux groupes de locuteurs non natifs nous amènent à présenter les systèmes vocaliques du Roumain et de l'Arabe du Maghreb afin de les comparer aux deux langues apprises : le Catalan et l'Espagnol. Il s'agit aussi de comprendre si ces divergences sont influencées par les propriétés de la L1.

1.2 Influence des systèmes vocaliques de L1 sur la production des contrastes non natifs

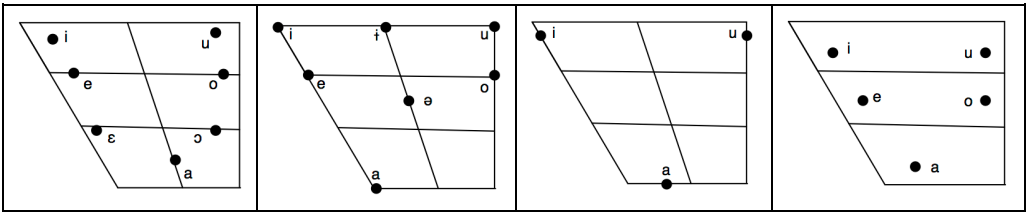


FIGURE 2 : Systèmes vocaliques des langues à l'étude (par ordre de gauche à droite) : Catalan région nord-est (adapté de Carbonell et Llisterri, 1999), Roumain (Chiţoran, 2001), Arabe du Maghreb (Thel Wall & Sa'Adeddin, 1999), Espagnol Castillan (Martinez-Celdran *et al.*, 2003).

Dans la perspective psycholinguistique traditionnelle, l'analyse de la similarité phonologique entre les différentes langues est considérée comme un outil permettant de prédire comment la L1 peut interférer dans la perception et la production d'une langue seconde (L2) (Strange, 2007). La figure 2 présente les systèmes vocaliques des langues étudiées. Du point de vue des origines, le catalan, l'espagnol et le roumain partagent des propriétés communes étant toutes trois des langues latines. Du point de vue phonologique, les voyelles au sein de chacun de ces systèmes s'opposent par les traits d'ouverture, d'antériorité et d'arrondissement labial. Là où le roumain se distingue par la présence de voyelles centrales, le catalan de Lleida se distingue par la présence de voyelles mi-ouvertes. Finalement, le système phonologique de l'arabe est le plus éloigné des trois autres par l'absence de voyelles moyennes et de voyelles centrales.

Ainsi, la difficulté à produire le contraste entre les voyelles moyennes du catalan pour les catalanophones non natifs (*cf.* Figure 1) pourrait s'expliquer par : a) l'absence de voyelles mi-ouvertes dans le système vocalique roumain et b) l'absence du contraste entre les voyelles mi-ouvertes et mi-fermées dans le système vocalique arabe (*cf.* Figure 2). Conformément aux théories qui traitent du rôle de la langue maternelle (L1) sur l'acquisition de contrastes non natifs (*cf.* Best, 1995, Kuhl & Iverson, 1995, Flege, 1995), on peut entrevoir dans les deux cas un transfert linguistique des propriétés de la L1 des locuteurs non natifs du CAT sur la production de la L2.

1.3 Objectif, questions et hypothèses de l'étude

Dans cette étude, nous nous interrogeons sur la réception, par des auditeurs natifs du catalan, des déviations de voyelles catalanes produites dans le contexte multilingue décrit auparavant. Pour répondre à cette problématique, nous proposons à des auditeurs catalanophones natifs un test de

catégorisation libre de ces productions. Ce type de test présente la particularité de ne restreindre ni le nombre d'écoutes des stimuli ni le temps de réalisation de la tâche.

Nous nous référons au modèle ASP (pour *Automatic Selective Perception* - Strange, 2011), pour formuler nos hypothèses. Dans sa présentation du modèle, l'auteur indique qu'en fonction de la tâche et de la complexité des stimuli perçus, les auditeurs d'une langue utilisent deux modes de traitement référant à deux types de connaissances linguistiques : un mode phonétique ou un mode phonologique. Respectivement aux deux modes de traitement cités, nous pouvons imaginer deux façons de percevoir les déviations : a) soit les auditeurs perçoivent les détails allophoniques et considèrent les exemplaires vocaliques non natifs non acceptables en termes phonologiques, b) soit les auditeurs catalanophones ignorent les variations phonétiques produites par les locuteurs non natifs, c'est-à-dire qu'ils perçoivent seulement l'information contrastive pertinente qui permet de discriminer entre les catégories phonologiques.

Les catégories de stimuli formées par les auditeurs natifs permettront de discuter des effets du transfert des propriétés de la L1 sur la production non native.

2 Méthode

2.1 Participants

Vingt-et-un étudiants de l'Université de Lleida (Femme=16; Homme=5), âgés entre 19 et 29 ans (moyenne d'âge : 21 ans), ont participé à cette expérience. Les critères d'inclusion étaient les suivants : tous étaient nés en Catalogne, ils étaient locuteurs natifs du Catalan et utilisaient cette langue de manière prédominante; ils étaient aussi bilingues Catalan-Espagnol. Aucun des participants ne présentait de problèmes d'audition ou de vue.

2.2 Stimuli

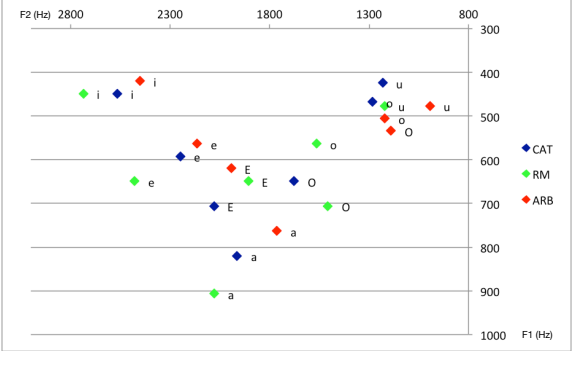


FIGURE 3 : Mesures formantiques (non normalisées) des voyelles catalanes sélectionnées parmi la base de données de Carrera et al. (2013). Ces voyelles sont produites par les locutrices dont la L1 est soit l'Arabe, soit le Roumain, soit le Catalan.

Au total, un ensemble de 21 logatomes (3 exemplaires de chacune des 7 voyelles catalanes) a été sélectionné parmi la base de données vocaliques constituée pour l'étude de Carrera-Sabaté (2013)

(*cf.* présentation en section 1.2). Au sein de chaque groupe linguistique CAT, ARB et RM, un exemplaire de chaque voyelle de l'inventaire vocalique du catalan /i e ɛ a o ɔ u/ a été sélectionné à l'oreille par deux phonéticiens catalanophones natifs. Le critère de sélection par les expérimentateurs était essentiellement perceptif, ceux-ci devant juger les exemplaires vocaliques s'approchant le plus d'une réalisation catalane.

Afin de limiter la variabilité des productions, les stimuli ont été sélectionnés parmi les logatomes de forme CVC avec C = [s]. La variabilité vocale a également été restreinte en sélectionnant les stimuli produits par le groupe féminin (20 filles au total avec CAT= 6 locutrices, ARB = 6 locutrices et RM = 8 locutrices). Pour autant, une importante variation subsiste. Cette variation est volontaire, car nos stimuli se veulent refléter la réalisation effective des voyelles par les différents groupes linguistiques étudiés. Les valeurs de F1-F2 des stimuli sélectionnés pour chaque groupe linguistique CAT, ARB et RM sont représentées en figure 3. Si nous comparons cette figure à la figure 1, une différence est notable concernant les valeurs, celles-ci diffèrent naturellement, car la figure 1 reporte des valeurs moyennes et normalisées de voyelles produites par des locuteurs tous genres confondus tandis que la figure 3 reporte les valeurs réelles et non normalisées des productions des locutrices (telles qu'elles ont été entendues par les auditeurs natifs dans la tâche de catégorisation proposée dans cette étude).

2.3 Procédure

Les participants ont effectué une tâche de catégorisation libre (TCL). La consigne était de grouper les stimuli en fonction des catégories de voyelles de leur langue. La tâche a été réalisée dans une salle calme du Laboratoire Pere Barnils à l'Université de Lleida. Les 21 stimuli sonores étaient présentés simultanément aux sujets via un écran d'ordinateur. L'interface informatique utilisée est TCL-LabX (http://petra.univ-tlse2.fr/tcl-labx/, Gaillard, 2009). Elle permet de présenter les stimuli visuellement (aléatoirement par des icônes numérotées) et auditivement, via un casque audio. Les participants pouvaient interagir de deux façons avec les icônes sonores : ils pouvaient écouter chaque stimulus en cliquant sur l'icône correspondant et former des groupes de stimuli semblables en les déplaçant sur l'écran. Les participants étaient informés que les stimuli mis dans un même groupe sont jugés similaires. Ils pouvaient écouter les stimuli autant de fois qu'ils le souhaitaient et dans l'ordre choisi. Ils pouvaient également faire autant de groupes qu'ils voulaient et n'étaient pas limités dans le temps pour la réalisation de la tâche.

2.4 Analyse des résultats

La méthode FAST - pour Factorial Approach for Sorting Task data - a été utilisée pour analyser les résultats (Cadoret *et coll.*, 2009). Cette méthode est basée sur une analyse à correspondance multiple (pour *Multiple Correspondance Analysis* ou MCA) permettant de proposer une représentation des objets et des catégories sur un plan à 2 dimensions (*cf.* Figure 4). Les deux dimensions (Axes 1 et 2 sur la figure 4) représentent les solutions les plus importantes pour la classification des stimuli (critères utilisés par les participants pour classer les sons). D'un point de vue statistique, le pourcentage associé à chaque dimension (soit 21.08% pour la dimension 1 et 19.09% pour la dimension 2) représente la contribution de cette dimension à la variance totale. Sur ce plan, chaque stimulus est représenté par un point. Deux stimuli sont d'autant plus proches qu'ils ont été classés ensemble par le plus grand nombre de participants.

Entre autres, l'intérêt de FAST est de complémenter les outils d'analyse offerts par la MCA en fournissant des éléments de validation grâce à une représentation en ellipses de confiance des

catégories moyennes formées par les sujets (*cf.* Figure 4). Pour le détail du calcul et de la méthode permettant d'obtenir ces représentations, nous renvoyons le lecteur à l'article de présentation de la méthode FAST sus-cité (Cadoret *et coll.,* 2009).

3 Résultats

3.1 Consensus entre les classes de 21 participants

Les résultats indiquent que les stimuli avec les voyelles [i], [ɛ] et [u] ont été groupés de façon quasi unanime (> 80.95%, c'est-à-dire plus de 17 fois). Les associations entre les stimuli avec la voyelle [a] sont aussi très récurrentes (en moyenne, les stimuli ont été classés ensemble par 76.19% des participants, c'est-à-dire 16 fois).

Ce sont les stimuli avec les voyelles [e, o, ɔ] qui présentent les scores d'association les plus faibles. Pour le groupe des voyelles [e], les scores diffèrent en fonction des associations observées : les productions CAT et RM de cette voyelle ont été associées de façon quasi unanime par les participants (19 fois) ; tandis que les associations entre les productions ARB et CAT d'une part, et les productions ARB et RM d'autre part, sont moins fréquentes (respectivement, 15 et 13 fois). On observe des scores plus hétérogènes encore concernant les associations entre les voyelles [o] et [ɔ]. Cette hétérogénéité montre que le classement de ces stimuli a davantage divisé les participants. Les associations avec d'autres catégories vocaliques sont aussi plus fréquentes pour ces deux voyelles. Par exemple, le détail des associations de la voyelle [o] avec les autres voyelles fait apparaître plusieurs associations avec la voyelle [ɔ], et dans une moindre mesure, avec la voyelle [u] (*ex.*: Ar-SoS a été associé 19 fois avec Ar-SɔS, 16 fois avec Ca-SoS, 11 fois avec Ca-SuS, 8 fois avec Ro-SoS, etc.).

3.2 Représentation des stimuli en fonction des critères de classification et stabilité des classes

La figure 4 met en évidence des groupes de stimuli très compacts pour les voyelles extrêmes [a, i, u] et la voyelle moyenne [ɛ]. Concernant les voyelles moyennes [e, o, ɔ], les distances sont plus importantes entre les stimuli au sein des groupes et d'autant plus pour la voyelle [e]. Les distances très proches entre les voyelles [o, ɔ] et parfois [u] montrent l'hétérogénéité des associations produites pour ces stimuli (*cf.* détail en section 3.1).

Les ellipses figurant sur le graphique représentent la variabilité interindividuelle des classes réalisées. Ainsi, on remarque une variation dans la taille des ellipses. On observe également des chevauchements importants entre les stimuli proches. Enfin, la différence majeure entre les ellipses tient dans leur forme. En fonction des variations sur les axes de dimensions 1 et 2, les ellipses s'étirent verticalement ou horizontalement. Pour les voyelles extrêmes [a, i, u], la forme des ellipses varie sur l'axe de dimension 1 (horizontale) tandis que pour les voyelles moyennes, la forme des ellipses varie sur l'axe de dimension 2 (verticale) (y compris pour la voyelle moyenne [ɛ] malgré le fort consensus sur son classement).

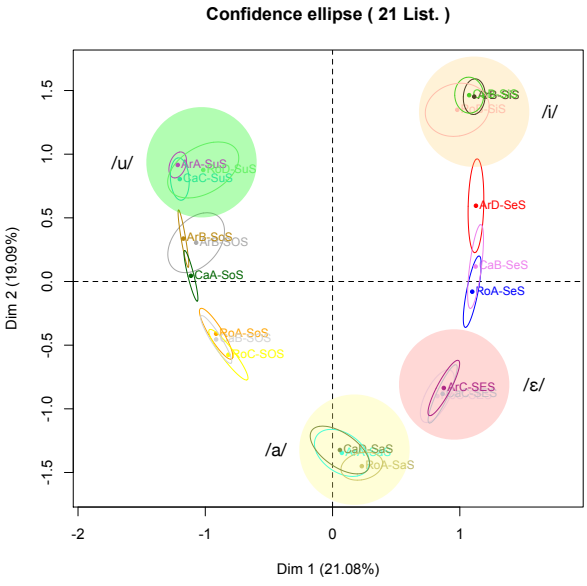


FIGURE 4 : Carte perceptuelle représentant les catégories formées par les auditeurs en fonction de deux dimensions. Les ellipses témoignent de la stabilité de ces catégories (Ar: Arabe/ Ca: Catalan / Ro: Roumain// Stimuli: S_S avec les voyelles: u, o, O [ɔ], a, E [ɛ], e, i).

4 Discussion

L'analyse des classes formées par les auditeurs catalanophones natifs avec les 21 stimuli produits par des locuteurs multilingues fait apparaître une catégorisation selon deux dimensions principales. Ces deux dimensions proposent une répartition des voyelles très semblables à la classification traditionnelle en fonction de critères acoustico-articulatoires (F1-F2 et/ou aperture/antériorité). Cela nous permet d'interpréter les deux dimensions en fonction de ces deux critères.

La catégorisation des stimuli fait apparaître six classes hétérogènes. Parmi ces classes, quatre présentent les mêmes caractéristiques : elles regroupent systématiquement les trois exemplaires d'une même catégorie vocalique, soit /i/, /u/, /a/ et /ɛ/, sachant que chacun de ces exemplaires est prononcé par une locutrice dont l'origine linguistique diffère (CAT, RM et ARB). Les distances très proches entre les exemplaires des quatre catégories (*cf.* Figure 4) suggèrent que les productions des locutrices non natives ont été perçues par les auditeurs natifs comme similaires aux productions des locutrices CAT. En se basant exclusivement sur le traitement perceptif de ces quatre classes vocaliques et en tenant compte du modèle de perception ASP (Strange, 2011), nous pensons que les auditeurs natifs du Catalan ont perçu les stimuli CAT, ARB et RM avec les voyelles [i, u, a, ɛ] sur un mode phonologique et ont ignoré les variations phonétiques dues à l'origine linguistique. Dans ce mode de traitement, tout se passe comme si les auditeurs re-formaient le système phonologique à partir des voyelles extrêmes /i, a, u/, en adéquation avec la *Quantal Theory of Speech* (Stevens, 1989) stipulant que ces voyelles sont les plus stables du point de vue des représentations.

En revanche les deux autres classes de stimuli avec les voyelles [o, ɔ] et [e] témoignent de difficultés relatives au traitement perceptif des exemplaires phonétiques qui les composent. Les distances plus importantes entre les stimuli au sein de ces deux groupes montrent en effet que leur catégorisation n'a pas fait l'objet d'un consensus entre les auditeurs CAT. La forme des ellipses qui varie sur l'axe de dimension 2, uniquement pour les voyelles moyennes, informe sur la nature des difficultés portant sur la distinction des stimuli en fonction de l'aperture (cette variation est également observée dans une moindre mesure pour la voyelle [ɛ]). Dans ce cas, il semble que la perception est corrélée avec les données acoustiques détaillées en production (*cf.* Carrera-Sabaté, 2013) et que les auditeurs ont catégorisé les voyelles moyennes selon un mode phonétique. Finalement, deux processus de traitement sont convoqués pour cette tâche proposant aux participants de catégoriser des stimuli variables et complexes en langue native. Nos deux hypothèses sont donc confirmées. En adéquation avec l'étude de Magnen & Gaillard, (2014) où les auteurs utilisaient également une tâche de catégorisation de stimuli complexes en langue native, nous observons toutefois que les deux processus peuvent coexister et être convoqués au sein de la même tâche.

Par ailleurs, les résultats de la tâche de catégorisation donnent des informations sur l'acquisition des catégories phonologiques non-natives par les locutrices multilingues. Pour les voyelles extrêmes qui n'ont pas posé de problèmes de classification aux auditeurs natifs, nous pouvons définir qu'elles ont été réalisées proches de catégories natives, car perçues comme telles. Ceci n'est pas vrai pour les voyelles moyennes. La forme et l'orientation des ellipses pour les groupes de voyelles moyennes (*cf*. Figure 4) tendent souvent vers des exemplaires proches des catégories existant dans la L1 des locuteurs non natifs. Par exemple, la réalisation de la voyelle moyenne [e] ARB a parfois été associée aux stimuli avec la voyelle [i], /i/ constituant la catégorie vocalique de l'arabe la plus proche de la catégorie vocalique du catalan /e/. Cet état de fait est cohérent avec les modèles d'assimilation de contrastes non natifs à des catégories natives (*cf.* Best, 1995).

5 Conclusion

La tâche de catégorisation de voyelles catalanes produites par des locutrices de diverses origines linguistiques a mis à jour différentes stratégies dans la classification de ces stimuli par les auditeurs natifs. Ces stratégies peuvent révéler deux modes de perception adaptés en fonction des difficultés posées par la variation phonétique des stimuli : un mode phonologique et un mode phonétique (Strange, 2011). Dans le mode de traitement phonologique, les légères variations phonétiques dans la production des voyelles extrêmes [i, a, u] par les différentes locutrices n'empêchent pas les auditeurs de recréer les catégories vocaliques de leur langue en incluant les exemplaires non natifs. En parallèle, les résultats observés dans le mode de perception phonétique mettent à jour une corrélation avec les données en production (articulatoires et acoustiques) et reflètent le transfert des propriétés phonologiques de la L1 vers les langues secondes (Best, 1995, Kuhl & Iverson, 1995, Flege, 1995)).

Remerciements

Cette étude a été réalisée dans le cadre d'un projet financé par MINECO (Spain) FFI2013-46987-C3. Nous remercions tout particulièrement le professeur Joan Julià-Muné ainsi que les professeurs et étudiants du Collegi Episcopal et de l'Institut Joan Oró de Lleida. Nous remercions également les étudiants de l'Université de Lleida.

Références

CARBONELL J., LLISTERRI J. (1999). Catalan. *Handbook of the International Phonetic Association.* Cambridge: Cambridge University Press, 62.

CARRERA-SABATE J. (2013). Vocals tòniques del lleidatà en entorns multilingües. *Treballs de Sociolingüística Catalana* 23, 117-132. Version electronique: http://publicacions.iec.cat/repository/pdf/00000205/00000072.pdf .

CHIȚORAN I. (2001). *The Phonology of Romanian: A Constraint-Based Approach*. Berlin et New York: Mouton de Gruyter.

CADORET M., LE S., PAGES J. (2009). A Factorial Approach for Sorting Task data (FAST). *Food Quality and Preference*, 20(6), 410–417.

FLEGE J. E. (1995). Second language speech learning: Theory, findings, and problems. W. Strange (ed.) *Speech perception and linguistic experience: Issues in cross-linguistic research*, Timonium, MD: York Press, 233-277.

GAILLARD P. (2009). Laissez-nous trier ! TCL-LabX et les tâches de catégorisation libre de sons. D. Dubois (ed.) *Le Sentir et le Dire : Concepts et méthodes en psychologie et linguistique cognitives*. Paris : L'harmattan, 189-210.

KUHL P. K., IVERSON, P. (1995). Linguistic Experience and the « Perceptual Magnet Effect. » W. Strange (ed.) *Speech Perception and Linguistic Experience*. Baltimore: York Press, 121–154.

LORES E., SOTO J., BERENGUER O. (2010). *Les altres llengües a Lleida. Mapa linguistic de Lleida*. Lleida: Òmnium Cultural et Pagès Editors. Version electronique http://lleida.omnium.cat; http://www.paeria.es/dcci.

MAGNEN C., GAILLARD P. (2014). Catégorisation de distorsions vocaliques produites par un apprenant hispanophone adulte en français L2. *Congrès Mondial de Linguistique Française, CMLF 2014*, Berlin, 1329-1343.

STEVENS K.N. (1989). On the quantal nature of speech. *Journal of Phonetics*, 17, 3-46.

STRANGE W. (2007). Cross-language phonetic similarity of vowels: Theoretical and methodological issues. O.-S. Bohn et J. M. Munro (ed.) In *Language Experience in Second Language Speech Learning: In honor of James Emil Flege*. Amsterdam: John Benjamins Publishing Company, 35-55.

STRANGE W. (2011). Automatic selective perception (ASP) of first and second language speech : A working model. *Journal of Phonetics*, 39(4), 456–466.

THEL WALL R., SA'ADEDDIN A. (1999). Arabic. *Handbook of the International PhoneticAssociation*. Cambridge: Cambridge University Press.

Peut-on caractériser globalement une « qualité d'acte expressif » : de « *breathy voice* » à « *breathy turn taking* » dans la glu socio-affective de l'interaction humain-robot ?

Liliya Tsvetanova Véronique Aubergé Yuko Sasa
Laboratoire Informatique de Grenoble – UMR5217, 38041 Saint Martin d'Hères, France
Prenom.Nom@imag.fr

RESUME

L'interaction face-à-face est considérée ici comme un système émergeant, englobant les sous-systèmes en synchronie des interactants inscrits, à travers leur personnalité, dans leur rôle social, leurs motivations, leurs intentions, leurs états socio-affectifs. L'interaction est instanciée par une « glu » socio-affective pour laquelle nous testons une dimension altruiste, orthogonale à la dimension de dominance, expérimentée dans le scénario écologique Emoz (Sasa et Aubergé, 2014) pour des personnes âgées donnant des commandes domotiques de forme imposée à un robot. Le dialogue est conduit par des *feedbacks* socio-affectifs primitifs du robot supposés « gluer » progressivement. Nous montrons que la variation faite par les sujets autour des commandes référentes, non seulement suit un décours dynamique de « glu » progressive, mais que le comportement communicatif des sujets est globalement inscrit dans des caractéristiques d'« intimité-*care*» d'une production *breathy* de toutes les modalités (voix, prosodie, paraphrasage lexico-morpho-syntaxique, timing, posture, direction du regard, proxémie, déplacement).

ABSTRACT

Multidimensional prosodic style, as characteristics of the "gluing" relation process: extension of "*breathiness*" from voice quality to "turn talk quality"

In this paper, the face-to-face interaction is considered as an emerging social system including two subsystems in synchrony in which the speakers are represented with their personality, social role, motivations, intentions and socio-emotional states. The interaction commence with a socio-affective « gluing » process which is experimented with the ecological scenario Emoz (Sasa et Aubergé, 2014) in which elderly appoint imposed domotic commands to a service robot called Emox. The dialogue is governed by the use of primitive socio-emotional robot feedbacks, considered to gradually simulate the « gluing » process. In this paper, we expose that not only the transformations of the commands given to Emox follow the socio-affective « gluing » dynamics, but also the elderly global communication behaviour can be characterised by the concept of « intimacy – care » reflected by the breathiness perceived in all communication modalities (voice, prosody, lexico-morpho-syntactical paraphrasing, timing, posture, gaze direction, proxemics and movements).

MOTS-CLES : dialogue socio-affectif, prosodie expressive, paralinguistique, paraphrasage, inter-synchronie, *breathiness*, multimodalité, Interaction Homme-Robot

KEYWORDS: socio-affective dialog, expressive prosody, paralinguistics, paraphrasing, inter-synchrony, breathiness, multimodality, Human-Robot Interaction

1 Introduction

L'interaction sociale est un processus dynamique, qui évolue dans le temps cognitif de la relation, dans lequel les interactants, amorcés *a priori* par la dyade de leurs rôles sociaux, co-construisent dialogiquement leur relation par les substrats de leurs personnalités dans leurs motivations, attentions mutuelles, processus socio-affectifs. L'étude présentée ici s'inscrit dans un travail éco-expérimental plus large, dédié à :

(1) contribuer à montrer qu'il existe une dimension altruiste, orthogonale à la dimension de dominance dans les socio-relations (*mutual social grooming*, Nelson and Geher, 2007) ;

(2) montrer que la relation est instanciée par une « glu » socio-affective dont le matériel primitif est la prosodie pure (Vanpé and Auberge, 2012);

(3) observer que la relation, se construisant dynamiquement (Delaherche and Chetouani, 2012 ; Morency, 2010), peut être conduite par des productions primitives, progressives et co-évolutives (Aubergé et al., 2013) ;

(4) montrer que la prosodie, incluant la qualité de voix (Campbell and Mokhtari, 2003) est un marqueur majeur des natures et des valeurs de la relation, mais que dans l'ensemble des modalités de production, la « manière » dont sont fabriquées les formes – forme des formes – révèle en cohérence ces mêmes natures et valeurs ;

(5) montrer qu'il existe une dimension altruiste de la relation qui est caractérisable par exemple par la co-construction gradiente d'une intimité—*care*, exprimée dans les productions multi-modales (Schröder et al., 2006) par le « style » cohérent de production des formes paralinguistiques (bruits de bouche), prosodiques, lexico-morpho-syntaxiques (paraphrasages autour des commandes imposées), posturales (incluant le regard (Argyle and Dean, 1965), mais pas les expressions faciales qui n'ont pas été recueillies proprement à cause de contraintes techniques), proxémiques statiques et dynamiques ;

(6) enfin, et c'est le but de notre présente étude, que toutes les modalités sont modifiées dans un style homologue, cohérent, de leurs formes : des changements au niveau de la qualité de voix des personnes âgées qui se résument dans une *breathiness* graduellement croissante et la gestualité corporelle qui joue également un rôle essentiel dans le processus multimodal interactionnel (Schröder et al., 2006). Dans une expérience – corpus GEE (Guillaume et al., 2015) - basée sur les mêmes primitives de construction de la « glu » pour des sujets jeunes empêchés de produire des expressions vocales, il a été montré que les gestes deviennent de plus en plus subtils quand l'intimité croît (et varient aussi avec la culture); les regards ont été montrés variants quand l'intimité augmente (Argyle and Dean, 1965) ; le déplacement s'adapte au rythme de l'autre (Schmidt, 2011), la proximité physique diminue (Scherer and Schiff, 1973).

Nous étudions donc ici comment les productions des sujets, dans plusieurs modalités, révèlent dans leurs formes le niveau croissant de la « glu », dans cette construction de dimension altruiste, pendant les interactions dans le scénario écologique Emoz, qui a permis de recueillir le corpus EEE (Elderly Emox Expressions) de dialogues spontanés entre des personnes âgées socialement isolées (afin de mieux mesurer le facteur de la « glu » pour un lien endommagé) et le robot Emox (développé par la société Awabot). Ce travail a également pour but technologique d'apporter au futur système de dialogue SADI (Socio Affective Dynamic Interactions) des évaluations expertes pour la supervision des modules évolutifs de *machine learning* de la reconnaissance automatique de la parole (RAP) et de la synthèse vocale.

2 Corpus EEE – recueil de dialogues spontanés « gluants »

Le corpus de dialogues spontanés multimodaux analysé dans cette étude – corpus EEE – a été collecté lors d'une expérience basée sur le principe d'une méthode d'expérimentation organisée selon une boucle agile de 3P (Public-Private-People) qui réunit tous les acteurs - académiques, industriels ou sociaux - dans le but de co-construire un processus d'expérimentation écologique en plaçant au centre l'humain, en faisant évoluer itérativement la boucle hypothèses/données comportementales, par une méthodologie qui oblige chaque acteur à assumer le devoir éthique attaché à son rôle. Les données du corpus EEE ont été collectées dans le Living Lab du LIG (Laboratoire Informatique de Grenoble) – Domus – recréant un contexte micro-écologique en utilisant un scénario (Sasa and Aubergé, 2014) du type Magicien d'Oz. Ce corpus est constitué d'interactions spontanées multimodales entre des personnes âgées socialement isolées et le robot Emox – majordome de l'appartement Domus, dont les sons sont des primitives langagières chargées d'informations graduellement organisées selon le niveau de « glu » socio-affective et envoyées en réponse des commandes domotiques (une liste de 31 commandes permettant de contrôler les actions de Domus) que le sujet donne à Emox. L'établissement d'un processus dynamique affectif se révèle difficile pour les personnes âgées en situation d'isolement social, sachant que leurs capacités nécessaires à la création de la « glu » socio-affective sont endommagées (Aubergé et al., 2014). Cette situation contrastive, favorisant l'observation de l'établissement d'une « glu » existante dans l'interaction entre le robot et la personne âgée, peut faire du robot un outil de réentrainement des personnes au processus dynamique de création et du maintien de la « glu » socio-relationnelle dans le but de faciliter la création du lien communicatif dans une interaction humain-humain.

L'ensemble des données du corpus est auto-annoté, via une annotation non experte basée sur une méthodologie rigoureuse inspirée de l'ethnométhodologie et permettant de réduire le biais de l'interprétation subjective et/ou gnostique. Lors d'une visualisation de leurs actes durant l'expérimentation, effectuée peu après l'expérimentation (environ 1 mois), les sujets font appel à leur mémoire autobiographique (Williams et al., 2007), ce qui permet de poser des étiquettes marqueurs de l'état affectif des personnes âgées au moment de l'expérience. Cette étape permet (1) de définir le niveau de « glu » socio-affective à chaque moment de l'expérience, (2) d'observer sa transformation globale et progressive, et également, (3) de détecter les frontières de chaque niveau de « glu » relationnelle.

Pour cette étude nous avons retenu 6 sujets féminins âgées de 67 à 89 ans dont le niveau de dépendance correspond au GIR5 et GIR6 (selon l'échelle de dépendance AGGIR). Les séquences audio-visuelles d'interaction entre chaque personne âgée et Emox, d'une durée de 40 à 53 minutes, ont été étiquetées en termes de comportements verbaux et non verbaux. Le panel observé est, pour le moment, de taille réduite car l'étiquetage descriptif des événements multimodaux hors parole s'effectue entièrement manuellement.

3 Marqueurs de la « glu » socio-affective

3.1 Valeurs externes d'auto-annotation

Les étiquettes d'auto-annotation permettent de regrouper l'ensemble des comportements dans des séquences globales démontrant l'évolution dynamique de l'état affectif du sujet dans le temps

dialogique. Néanmoins, bien que le changement de la relation avec le robot soit relevé pour l'ensemble des sujets, l'auto-annotation étant basée sur la réminiscence des processus cognitifs pas à pas sur les temps interactionnels, les étiquettes obtenues sont très différentes d'un individu à l'autre. Un exemple typique de progression graduelle de la « glu » est l'étiquette du matériel langagier (commande) comme « ordre » qui devient au fur et à mesure de l'avancement de l'expérience « ordre gentil / ordre en collaboration », ensuite « interrogation », « demande », « suggestion » ou bien l'étiquette « ton impératif » utilisée au début de l'expérience qui devient « ton familial ».

3.2 Paraphrasage lexico-morpho-syntaxique

Rappelons que la consigne donnée aux sujets est de respecter scrupuleusement l'énoncé figé pour chaque commande. Or, l'une des observations importantes du corpus EEE est l'apparition de paraphrasage au cours de l'expérience. L'analyse morpho-syntaxique et lexicale des paraphrases a mis en évidence des changements qui interviennent de manière graduelle mais systématiquement ordonnée en suivant la courbe évolutive de la « glu » relationnelle. Pour pouvoir déterminer de quelle manière s'effectue le paraphrasage des commandes au fur et à mesure de l'expérience, nous avons procédé par une observation des changements langagiers (morphologiques, lexicaux et syntaxiques) par rapport à la commande de référence qui est notée sur la feuille des commandes imposée aux sujets au début de l'expérience, en sachant que toutes les 31 commandes domotiques sont composées d'un infinitif verbal suivi d'un constituant nominal (ex. « Mettre la lumière »). La variation de la forme langagière est significative pour le niveau de « glu », puisqu'elle est surtout observable chez les sujets qui ont le plus « glué » (guidé par les auto-annotations) avec Emox et, en plus, la variation des formes langagières suit exactement la même tendance d'évolution pour tous les sujets. Comme montré par la figure 1, l'infinitif de la commande de référence est substitué progressivement par des formes conjuguées indiquant des changements au niveau de la prise en considération du robot (d'abord par l'inclusion du robot dans un processus commun par l'utilisation du pronom personnel « on » et ensuite par la distinction du robot comme une entité par le pronom personnel « tu ») qui est fortement liée à la force de la « glu » socio-affective établie dans le temps cognitif dialogique (T) de l'expérience.

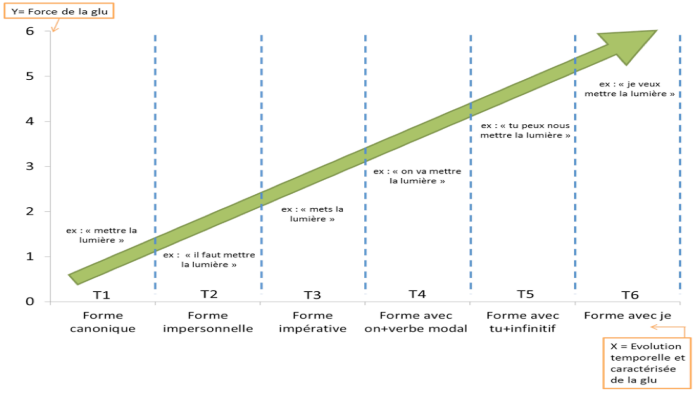


Figure 1: Evolution temporelle des transformations morphologiques, lexicales et syntaxiques par rapport au niveau de "glu" socio-affective défini par les valeurs externes d'auto-annotation

3.3 Evolution temporelle de la « *breathiness* »

La qualité de voix est une composante prosodique qui transmet des informations sur l'état affectif du locuteur et la *breathiness* est un type de qualité de voix spécifiquement lié à l'existence d'un lien intime entre les interactants. D'un point de vue articulatoire, la *breathiness*, qui est acoustiquement associée à un effet d'intimité, est produite par un relâchement des cordes vocales suite à une tension musculaire faible, mais selon le niveau de faiblesse musculaire, la *breathiness* acoustique peut être perçue comme plus *tense* (tendue) ou plus *lax* (détendue). La *breathiness* dans la voix âgée en situation de parole spontanée est très difficile à évaluer, car les changements physiologiques liés à l'âge ont une influence sur le relâchement de la tension musculaire qui, par conséquent, affecte les muscles du larynx induisant ainsi une perception acoustique de la voix âgée comme étant plus *breathy*. Ainsi, les mesures traditionnellement utilisées pour détecter la *breathiness* (H1H2, Harmonics-to-Noise Ratio, Glottal-to-Noise ratio, etc.) n'étant pas adaptées à la voix âgée, nous avons procédé à un étiquetage « expert » de la *breathiness* acoustique dont les étiquettes sont graduellement organisées par rapport au niveau de *breathiness* perçue (figure 2). Comme montré par la figure 2, la *breathiness* des personnes âgées est présente dès que la « glu » s'installe, mais elle évolue subtilement selon la tension de plus *tense* à plus *lax* en parallèle avec le niveau de « glu » défini par les étiquettes des auto-annotations.

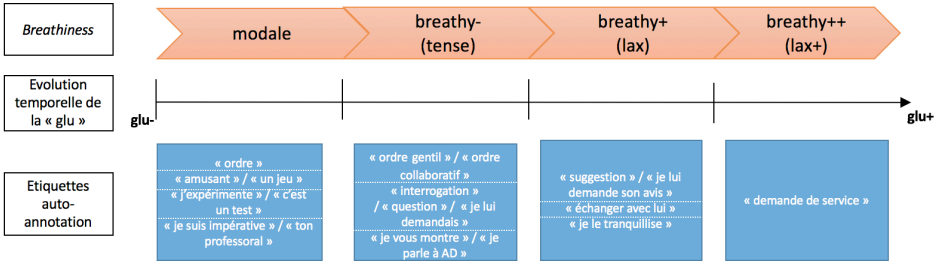


Figure 2: Evolution de la breathiness dans le corpus EEEselon l'axe de la « glu » socio-relationnelle (cf. Figure 1)

L'évolution de la *breathiness* dans le corpus EEE est également marquée par des changements ponctuels et prompts du niveau de *breathiness* qui paraissent être sémantiquement importants pour le niveau de « glu » dans lequel la relation entre le sujet et Emox peut être située. Ainsi, si le robot ne réalise pas la commande demandée avant un certain seuil de « glu », si par exemple l'erreur intervient au début de l'expérience, le niveau (n) de *breathiness* acoustique dans la commande que réitère le sujet pour faire exécuter la commande correctement diminue (n-1) ou souvent la voix redevient modale, voire même le niveau de *breathiness* devient négatif (dans les cas d'irritation, d'impatience) et il est alors plus long de ré-intéresser/« re-gluer » le sujet (même si le sujet a pour consigne obligatoire de finir d'énoncer l'ensemble des commandes). Lorsque l'erreur d'exécution du robot intervient après un seuil reconnaissable par l'apparition de valeurs des marqueurs qui sont décrits ici – seuil dont nous voulons montrer plus tard qu'il est un changement catégoriel dans le continuum de l'attachement – alors au contraire le sujet lorsqu'il réitère la commande, exprime une progression forte de la « glu » (*breathy* (n+i)), et inhabituelle dans le reste de son comportement : à ce stade, plus le robot se trompe, plus le sujet s'attache en l'aidant à récupérer son erreur.

4 Dynamique des modalités interactionnelles non verbales

4.1 Evolution temporelle des modalités (selon la « glu » socio-affective)

L'initiation et la progression de la « glu » relationnelle s'inscrivent dans une dynamique strictement individuelle, mais, bien que les événements multimodaux se produisent à des niveaux temporels différents selon les caractéristiques propres de chaque sujet et selon la progression de la « glu », ils sont observables et similaires chez tous les sujets de l'étude à un niveau de « glu » équivalent. Ainsi, les personnes qui ont moins « glué » avec le robot reflètent les mêmes comportements de gestualité corporelle multimodale (direction du regard, déplacement, posture et proxémie) que ceux que les sujets qui ont le plus « glué » avec Emox effectuent au début de leurs expériences.

La direction du regard des sujets est orientée sur 4 cibles principales (environnement, objet de la commande, feuille de commandes, Emox), chacune ayant une dynamique de progression différente. Par exemple, toutes les personnes âgées regardent l'environnement et l'objet de la commande le plus souvent au début de l'expérience et la fréquence de fixation de ces cibles diminue avec l'habituation à l'environnement. Les regards en direction d'Emox sont variants, mais il a été constaté que les personnes ayant moins « glué » avec Emox le regardent moins, alors que pour l'autre groupe de personnes âgées, en termes de durée de fixation, les regards en direction d'Emox sont généralement moins longs, mais plus fréquents en fin d'expérience (où le niveau de « glu » est plus élevé). Lorsqu'une réitération de commande se produit, les sujets fixent le plus souvent Emox. Quant aux regards en direction de la feuille, qui est un artefact imposé aux personnes âgées leur permettant d'adresser des commandes à Emox, il a été observé un détachement progressif, aussi bien physique (la personne ne tient pas la feuille et ne la regarde pas) que psychologique (la personne regarde la feuille, mais modifie la forme langagière des commandes inscrite sur la feuille), chez les personnes âgées ayant le plus « glué » avec Emox, alors que les personnes âgées qui n'arrivent pas tout à fait à rentrer dans une relation affective avec Emox, se maintiennent, physiquement et psychologiquement, à la feuille (regard fréquent en direction de la feuille lors de l'énonciation des commandes accentué par une rare variation de la forme langagière).

Les premiers déplacements des sujets dans l'appartement sont motivés par le scénario, parce que les participants sont incités à appliquer l'ensemble des commandes dans chaque pièce, mais il a été observé que toutes les personnes âgées ont tendance à marcher devant Emox, mais lorsqu'elles s'arrêtent à l'endroit cible de leur action de déplacement, elles attendent la fin du déplacement d'Emox avant de lui donner une nouvelle commande. Ce mode de déplacement se reproduit tout le long de l'expérience pour les personnes ayant moins « glué » avec le robot, mais une différence est observée pour les personnes âgées qui ont créé une relation socio-affective avec Emox qui au fur et à mesure de l'expérience commencent à synchroniser leur manière de se déplacer par rapport à Emox : bien qu'Emox, étant petit, ne se déplace pas rapidement, les sujets marchent derrière le robot ou à côté de lui en l'attendant avancer.

La proxémie entre le sujet et Emox est naturellement adoptée par chaque sujet au début de l'expérience et si Emox réduit cette distance sociale choisie de manière spontanée par les personnes âgées, avant que la « glu » soit instaurée, les sujets reculent pour revenir à la distance initiale spontanément instaurée. Au fur et à mesure de l'expérience, les personnes ayant construit de la « glu » se rapprochent ou laissent Emox s'approcher d'eux de plus près et pour plus long temps comparé aux sujets ayant moins « glué ». Une autre observation relève qu'une fois la «

glu » instaurée, lorsque la distance physique entre un sujet et le robot est agrandie, les sujets ont tendance à fixer le plus souvent Emox.

La posture redressée et fléchie au début de l'expérience s'avachie au fur et à mesure et les premières apparitions de *breathiness* acoustique semblent correspondre souvent à une posture assise. Après l'établissement de la « glu », indépendamment du fait si les sujets sont debout ou assis, il a été observé que les sujets se penchent en avant vers Emox lorsqu'ils lui parlent ou lui donnent des commandes et l'émergence de ces penchements semble être relative avec une voix *breathy* très détendue (lax+) et une direction du regard vers Emox.

4.2 Evénements ponctuels observés significatif pour le niveau de « glu » socio-affective

4.2.1 Marquage par un focus prosodique

Certains sujets commencent à produire, à un certain niveau de « glu » établi, des focus prosodiques dont les accents acoustiques s'accompagnent d'un accent gestuel (mouvements rapides de la tête). Ces hochements de la tête en début d'expérience sont ciblés sur Emox et ont une dynamique plutôt rapide (parallèlement à des focus à voix *tense*), alors qu'à la fin de l'expérience, ils sont faits dans la direction de l'objet de la commande, mais sont aussi plus lents et relâchés quand la voix est *breathy* (*lax+*).

4.2.2 Apparition de comportements à dynamique évolutive

A un niveau de « glu » élevé, des phénomènes spécifiques apparaissent et leur dynamique semble aller de paire avec le niveau de *breathiness* et l'aspect *tense* et *lax* de la voix. Un de ces phénomènes est la gestualité et bien qu'elle fasse partie intégrante de l'interaction non verbale et comme montré par (Guillaume et al., 2015) a une signification par rapport au niveau de « glu », elle n'est pas un point expressément observé dans cette étude à cause du nombre limité de sujets. Néanmoins, des gestes très subtils - de pointage de l'objet de la commande, d'ouverture des paumes des deux mains quand le sujet est en train de s'adresser à Emox (hors commandes) – apparaissent et sont notables chez tous les sujets en fin d'expérience quand le niveau de « glu » est très élevé et la voix des sujets est très *breathy (lax+)*. D'autre part, bien que ponctuelles, certaines modalités semblent être co-occurrentes à un niveau de « glu » élevé, ce qui révèle une dynamique des changements. Par exemple, les penchements ou bien l'orientation du corps vers Emox lorsque le sujet a un fort lien affectif en fin d'expérience apparaissent toujours parallèlement à un regard fixé sur Emox et une qualité de voix *breathy lax*. Cette observation suggère l'existence d'une synergie des modalités qui pourrait donner une information subtile similaire à celle qui est caractérisée par la nature de la *breathiness* de la voix.

5 Discussion

La *breathiness* acoustique des sujets âgés faisant partie du corpus EEE est observée dès l'initiation des primitives langagières d'Emox qui sont graduellement chargées de « glu », ce qui, comme montré par l'auto-annotation, fait augmenter de manière parallèle et graduelle le niveau de « glu »

chez la personne âgée, et la *breathiness* instaurée ainsi n'évolue qu'au sein d'elle-même, dans sa nature, selon le degré de tension musculaire attribué allant de *tense* à *lax*. Les observations sur la *breathiness* acoustique laissent supposer qu'il existerait deux niveaux selon la vitesse d'évolution de la *breathiness* dans le corpus que nous pouvons caractériser d'une part comme « *breathiness* globale » qui s'étale sur toute la durée de l'expérience augmentant au fur et à mesure, et d'autre part, comme « *breathiness* locale » qui est observable ponctuellement à des moments précis modifiant inopinément la construction relationnelle. Ainsi, au sein de la « *breathiness* globale », plus le niveau de « glu » augmente et plus le niveau de *breathiness* audible augmente en diminuant parallèlement et subtilement la tension musculaire pour venir progressivement à une qualité de voix perceptivement détendue (*lax+*). Cette évolution globale et subtile est non seulement audible via la *breathiness*, mais elle semble également s'illustrer à travers le comportement global (expression corporelle) des sujets, car les observations sur certaines modalités interactionnelles non verbales ont révélé qu'elles évoluent de manière continue, mais différente selon les autres modalités et ce n'est pas le type au sein de la modalité qui change, mais la subtilité mise en ouvre pour la production d'une expression corporelle au sein d'un même type de comportement non verbal. Ce changement subtil dans l'expression corporelle, en lien avec le niveau de *breathiness* et très marqué en fin d'expérience, transparait dans la variation de la durée d'un même comportement (ex. diminution progressive du temps de fixation d'Emox), mais également dans l'adaptation par rapport au rythme de déplacement d'Emox et la réduction progressive de distance physique avec Emox. De plus, à un niveau de « *breathiness* locale », d'autres comportements non verbaux semblent marquer ponctuellement l'évolution de la « *breathiness* globale » aux moments de « glu » relationnelle très rigide. Ainsi, l'apparition d'une gestualité subtile et la tension diminuante observée dans la posture des sujets semblent être en lien avec la nature de la *breathiness* (à un niveau local) marquée à la fin de l'expérience par plusieurs moments ponctuels de *lax*, ce qui laisse supposer que la nature de la *breathiness* est observable dans le comportement global corporel des sujets, entre *tense* (tendu) et *lax* (détendu). L'ensemble des phénomènes d'expression corporelle font donc preuve d'une intimité instaurée progressivement chez les personnes âgées et accentuée à des moments précis de la progression de l'expérience et cette intimité correspond dans le déroulement temporel de l'expérience à un niveau de « glu » élevé justifié par le niveau de *breathiness* identifié dans la voix et détecté au même niveau temporel de l'expérience, mais également visuellement détectable.

6 Conclusion

La *breathiness* a été présentée dans cette étude comme une dimension dynamique des comportements multimodaux observés à travers des modalités verbales et non-verbales du processus interactionnel. Tous les phénomènes observés dans l'interaction entre les personnes âgées et le robot, qu'ils soient verbaux (nature de la *breathiness*, mais également la transformation des commandes) ou bien non verbaux (expressions corporelles) pris en considération dans cette étude montrent que les personnes âgées en interaction avec Emox adoptent un comportement globalement plus détendu et plus spontané, et donc plus intime envers Emox, ce qui laisse supposer qu'il existerait une qualité de *breathiness* plus globale, dépendante des niveaux de « glu » (intimité), marquant l'ensemble de l'interaction et perceptible acoustiquement et visuellement. La « glu » socio-relationnelle entre les participants dans une interaction dyadique peut donc être considérée comme un comportement plus global qui est plus complexe et plus riche dans la parole et l'implémentation langagière, mais qui est exprimé et perçu par la connaissance sensorimotrice de chaque personne.

Remerciements

Cette étude a été partiellement financée par la bourse Interabot (Projet Investissements d'Avenir) et a été partiellement supportée par le LabEx PERSYVAL-Lab (ANR -11-LABX 0025-01). Nous tenons à remercier la société Awabot, l'entreprise Bien A la Maison (services d'accompagnement des personnes âgées) et le Foyer logements de Gières pour leur collaboration active. Nous remercions également Romain Magnani, Frédéric Aman, Natasha Borel, Clarisse Bayol, Maxence Girard-Rivier, Nicolas Bonnefond pour leur participation au travail dans lequel cette étude n'est qu'un aspect.

Références

Argyle, M., and Dean, J. (1965). Eye-contact, distance and affiliation. Sociometry 289–304.

Aubergé, V., Sasa, Y., Robert, T., Bonnefond, N., and Meillon, B. (2013). Emoz: a wizard of Oz for emerging the socio-affective glue with a non humanoid companion robot. (Grenoble, France),.

Aubergé, V., Sasa, Y., Bonnefond, N., Meillon, B., Robert, T., Rey-Gorrez, J., Schwartz, A., Antunes, L.B., De Biasi, G., Caffiau, S., et al. (2014). >The EEE corpus: socio-affective" glue" cues in elderly-robot interactions in a Smart Home with the EmOz platform. In 5th International Workshop on EMOTION, SOCIAL SIGNALS, SENTIMENT & LINKED OPEN DATA, (Reykjavik, Iceland),.

Campbell, N., and Mokhtari, P. (2003). >Voice quality: the 4th prosodic dimension. In 15th ICPhS, (Barcelona, Spain), pp. pp. 2417–2420.

Delaherche, E., and Chetouani, M. (2012). Synchronie interpersonnelle: un panorama des méthodes d'évaluation. In WACAI 2012 Workshop Affect, Compagnon Artificiel, Interaction, pp. pp. 151–158.

Guillaume, L., Aubergé, V., Magnani, R., Aman, F., Cottier, C., Sasa, Y., Wolf, C., Nebout, F., Neverova, N., Bonnefond, N., et al. (2015). HRI in an ecological dynamic experiment: the GEE corpus based approach for the Emox robot. (Lyon, France),.

Morency, L.-P. (2010). Modeling human communication dynamics. IEEE Signal Process. Mag. *27*, 112–116.

Nelson, H., and Geher, G. (2007). Mutual grooming in human dyadic relationships: an ethological perspective. Curr. Psychol. *26*, 121–140.

Sasa, Y., and Aubergé, V. (2014). >Socio-affective interactions between a companion robot and elderly in a Smart Home context: prosody as the main vector of the "socio-affective glue." In Speech Prosody 2014, (Dublin, Ireland),.

Scherer, S.E., and Schiff, M.R. (1973). Perceived intimacy, physical distance and eye contact. Percept. Mot. Skills *36*, 835–841.

Schmidt, P.F. (2011). Understanding social motor coordination. Hum Mov Sci. Hum. Mov. Sci. *30*, 834–845.

Schröder, M., Heylen, D.K.J., and Poggi, I. (2006). Perception of non-verbal emotional listener feedback.

Vanpé, A., and Auberge, V. (2012). Early meaning before the phonemes concatenation? Prosodic cues for Feeling of Thinking. In GSCP, (Belo Horizonte, Brazil), pp. x – x.

Williams, H.L., Conway, M.A., and Cohen, G. (2007). 2 Autobiographical memory. Mem. Real World 21.

Phonétisation statistique adaptable d'énoncés pour le français

Gwénolé Lecorvé Damien Lolive
IRISA, Université de Rennes 1, Lannion, France
gwenole.lecorve@irisa.fr, damien.lolive@irisa.fr

RÉSUMÉ

Les méthodes classiques de phonétisation d'énoncés concatènent les prononciations hors-contexte des mots. Ce type d'approches est trop faible pour certaines langues, comme le français, où les transitions entre les mots impliquent des modifications de prononciation. De plus, cela rend difficile la modélisation de stratégies de prononciation globales, par exemple pour modéliser un locuteur ou un accent particulier. Pour palier ces problèmes, ce papier présente une approche originale pour la phonétisation du français afin de générer des variantes de prononciation dans le cas d'énoncés. Par l'emploi de champs aléatoires conditionnels et de transducteurs finis pondérés, cette approche propose un cadre statistique particulièrement souple et adaptable. Cette approche est évaluée sur un corpus de mots isolés et sur un corpus d'énoncés prononcés.

ABSTRACT

Adaptive statistical utterance phonetization for French *.

Traditional utterance phonetization methods concatenate pronunciations of uncontextualized constituent words. This approach is too weak for some languages, like French, where transitions between words imply pronunciation modifications. Moreover, it makes it difficult to consider global pronunciation strategies, for instance to model a specific speaker or a specific accent. To overcome these problems, this paper presents a new original phonetization approach for French to generate pronunciation variants of utterances. This approach offers a statistical and highly adaptive framework by relying on conditional random fields and weighted finite state transducers. The approach is evaluated on a corpus of isolated words and a corpus of spoken utterances.

MOTS-CLÉS : Phonétisation, variantes de prononciation, treillis de phonèmes, champs aléatoires conditionnels, transducteurs finis pondérés.

KEYWORDS: Utterance phonetization, pronunciation variant modelling, phoneme lattices, conditional random fields, weighted finite state transducers.

1 Introduction

La phonétisation a pour objectif de prédire une séquence de phonèmes à partir d'une séquence de graphèmes. Pour la plupart des langues, cette tâche se limite au cas de mots isolés, réduisant la phonétisation d'un énoncé aux prononciations concaténées de ses mots. Cette approche n'est cependant pas viable pour certaines langues, comme le français, où les transitions entre mots provoquent des modifications de leur prononciation, à moins d'inclure des informations, souvent minimales, sur le contexte phonologique. De plus, cette approche complique la modélisation de stratégies de prononciation globales, par ex. propres à un locuteur ou à un accent particulier. Cette tâche d'adaptation est

*. Cet article reprend un travail présenté par les mêmes auteurs à la conférence ICASSP 2015.

majeure, en particulier en synthèse de la parole (TTS) (Benesty *et al.*, 2008).

Pour palier ces problèmes, ce papier présente une nouvelle méthode pour la phonétisation du français. Cette méthode apporte trois contributions : (i) elle introduit la notion de modèle d'élision pour modéliser les variantes intra-mots ; (ii) elle intègre les contextes phonologiques pour modéliser les variantes inter-mots ; (iii) elle permet de générer des treillis probabilistes de phonèmes à partir d'énoncés, et non seulement de mots isolés. Pour cela, cette méthode repose sur des Champs Aléatoires Conditionnels (CAC) pour estimer les probabilités des phonèmes sur les mots isolés, puis sur des transducteurs finis pondérés (TFP) pour traiter les transitions entre mots. On obtient ainsi des treillis de phonèmes à partir desquels des phonétisations peuvent être dérivées.

Le potentiel de ce phonétiseur est très important. Les treillis de phonèmes générés offrent beaucoup de flexibilité puisque les transitions peuvent être repondérées en utilisant différents modèles de prononciation dédiés à une tâche donnée. Néanmoins, l'objectif de cet article est de présenter la méthode de phonétisation et de ses premiers résultats sans adaptation, et non d'étudier le caractère adaptable des treillis, ce dernier point étant conservé pour de futurs travaux. De manière plus générale, l'objectif de cet article n'est pas la recherche de résultats meilleurs que l'état de l'art mais plutôt de définir un cadre de travail générique et de démontrer son applicabilité pour le français. Ce cadre de travail peut être étendu aisément et complété avec de nouveaux modèles. De plus, il ne repose pas sur des règles expertes et peut donc facilement être porté à de nouvelles langues. Enfin, l'approche proposée peut également tolérer une certaine incertitude dans l'énoncé d'entrée, par exemple pour gérer plusieurs tokénisations.

Dans cet article, la section 2 présente le domaine, la section 3 et 4 introduisent notre méthode de phonétisation, respectivement pour des mots isolés, puis des énoncés. Les expériences y sont présentées sur le lexique de prononciation MHATLex et sur un corpus de parole.

2 État de l'art

La phonétisation est largement étudiée depuis des années, en particulier en reconnaissance automatique de la parole (RAP) et en TTS. La plupart des systèmes reposent principalement sur des lexiques de prononciation construits manuellement pour les mots communs avec une conversion graphème-phonème automatique pour les mots hors vocabulaire, c.-à-d. les mots qui ne sont pas dans le lexique. De nombreuses stratégies ont été proposées pour la conversion graphème-phonème dans la littérature : des méthodes à bases de règles (Béchet, 2001; Claveau, 2009), des approches statistiques (Bisani & Ney, 2008; Illina *et al.*, 2011), ainsi que d'autres approches variées (Bellegarda, 2005; Laurent *et al.*, 2009). Parmi celles-ci, les approches statistiques ont récemment montré des performances intéressantes tout en apportant la possibilité d'interpréter et d'adapter les scores des prononciations générées. Trois principales méthodes s'opposent. D'un côté, les méthodes à n-grammes joints reposent sur des séquences de paires graphème-phonème dont les probabilités sont habituellement obtenues avec des modèles de langage (Bisani & Ney, 2008; Novak *et al.*, 2012; Hahn *et al.*, 2012). Les CAC ont également prouvé leur performance pour traiter le problème de conversion graphème-phonème (Illina *et al.*, 2011; Wang & King, 2011; Hahn *et al.*, 2011; Lehnen *et al.*, 2012). Enfin, des méthodes fondées sur des réseaux de neurones ont également été proposées très récemment et semblent produire les meilleurs résultats (Rao *et al.*, 2015; Yao & Zweig, 2015). Dans cet article, nous nous appuyons sur les CAC car ils sont un moyen simple d'intégrer de multiples connaissances. Le portage de notre méthode à des réseaux de neurones est néanmoins parfaitement envisageable. Dans l'ensemble, nos travaux se rapprochent de (Illina *et al.*, 2011) pour la phonétisation de mots isolés, bien que des différences dans les protocoles expérimentaux empêchent des comparaions directes des résultats.

La prononciation des énoncés, c.-à-d. des séquences de mots, a été étudiée de manière plus partielle. En RAP, l'introduction de TFP comme moyen de décoder les signaux de parole a apporté une nouvelle représentation des alternatives de prononciation des mots (Mohri *et al.*, 2000). En particulier, (Hazen *et al.*, 2005) propose de représenter les énoncés, les prononciations et leurs variations possibles comme des TFP qui peuvent être composés et parcourus pour extraire des variantes de prononciation. Pour la conversion graphème-phonème de mots isolés, des TFP et des treillis de phonèmes (Bodenstab & Fanty, 2007; Polyákova & Bonafonte, 2011) ou encore des CAC (Lehnen *et al.*, 2011) ont également été utilisés pour représenter les alternatives. La philosophie de cet article est très proche en combinant CAC et TFP. Cependant, le travail présenté ici diffère de (Hazen *et al.*, 2005) puisque les mots hors-vocabulaire et les élisions sont traités ici. De plus, (Hazen *et al.*, 2005) se concentre sur l'anglais tandis que notre travail est réalisé sur le français qui est phonologiquement différent. La phonétisation des mots isolés est présentée dans la section 3 avant de passer au niveau énoncé en section 4.

3 Phonétisation des mots isolés

La phonétisation des mots isolés consiste à prédire une séquence de phonèmes à partir d'une séquence de graphèmes. Cette tâche peut être vue comme un problème d'étiquetage automatique. Dans ce travail, cet étiquetage est effectué par deux CAC employés consécutivement, l'un pour prédire une séquence de phonèmes, l'autre pour prédire d'éventuelles élisions sur ceux-ci. Cette section présente l'apprentissage de ces deux modèles, puis les expériences sur la phonétisation de mots isolés.

3.1 Modèle de phonétisation

Le problème de conversion graphème-phonème est traité par l'apprentissage d'un CAC, dit ici *modèle de phonétisation*, sur un corpus aligné de graphèmes et de phonèmes issus d'un lexique de prononciations. Pour de meilleures performances, comme souvent dans la littérature (Jiampojamarn *et al.*, 2007), ces alignements sont effectués entre blocs de graphèmes et phonèmes de taille maximale fixée. Suite à une étude préliminaire, ces blocs sont d'une taille maximale de 2 dans notre travail. Pour rendre possible la tâche d'étiquetage, ces blocs sont ensuite décomposés de telle sorte que tout phonème soit associé à un et un seul graphème et que chaque graphème soit associé à 0, 1 ou 2 phonèmes. Par exemple, les graphèmes « o n » peuvent être alignés au seul phonème /õ/ et « x » au bloc /ks/. Les graphèmes « o », « n » et « x » seraient alors respectivement associés à /õ/, /_/ (c.-à-d. aucun phonème) et /ks/. La littérature montre également qu'il est bon de compléter un graphème par ses voisins (Illina *et al.*, 2011; Wang & King, 2011). Ce voisinage est défini par une fenêtre de $\pm N$ graphèmes dont nous étudierons l'impact à la section 3.3. Enfin, comme le français contient de nombreux homographes aux prononciations différentes [1], la classe grammaticale est une autre information utile. Suivant le choix fait dans (Illina *et al.*, 2011), nous utilisons cette information en la simplifiant à la seule distinction verbe/non verbe. Sur le fond, d'autres caractéristiques comme l'étymologie du mot ou un dialecte à considérer pourraient être utilisées mais ce n'est pas l'objectif du présent travail. Formellement, le CAC que nous entraînons prédit donc chaque phonème p à partir d'un n-gramme de graphèmes $\mathbf{g}$ (le graphème associé et ses voisins) et d'autres caractéristiques $\mathbf{o}$ dérivées du mot à phonétiser. Ce CAC est capable de produire la ou les meilleures hypothèses de phonétisation du mot et de fournir la probabilité *a posteriori* $\phi(p|\mathbf{g}, \mathbf{o})$ de chaque phonème.

1. Par exemple, la graphie *président* se prononce /pʁezidɑ̃/ s'il s'agit du nom ou /pʁezid/ s'il s'agit du verbe *présider*.

3.2 Modèle d'élision

Comme dans d'autres langues, certains phonèmes du français peuvent être élidés. Ces élisions dépendent d'informations variées, comme le contexte phonologique, le type de parole, les règles ou exceptions liées à la grammaires, etc. Le phénomène le plus courant pour illustrer cette variabilité est le cas du schwa (/ə/) qui peut être élidé la plupart du temps. Par exemple, le mot *semaine* peut être prononcé /səmɛn/ ou /smɛn/. De plus, le dernier graphème *e* peut également être prononcé /ə/ lorsqu'il est suivi par une consonne. Ainsi, l'énoncé "*la semaine finit*" peut être prononcé /lasəmɛnfini/ ou /lasəmɛnəfini/. Dans certains contextes, la prononciation de schwas en position finale est une règle, par ex. en poésie. Cependant, toutes les occurrences de schwas ne sont pas optionnelles. Par exemple, le mot *Bretagne* est toujours prononcé /bʁətaɲ/. Des phénomènes similaires existent pour d'autres phonèmes, en particulier les liaisons lorsque l'on considère les liens entre mots consécutifs.

Dans cet article, nous proposons d'entraîner un autre CAC, dit *modèle d'élision*, pour prédire les élisions de phonèmes. Pour chaque phonème dans la prononciation d'un mot, l'étiquette à apprendre est soit *optionnel*, c.-à-d. que le phonème peut être prononcé ou non, soit *obligatoire*. En plus des graphèmes $\mathbf{g}$ et des autres caractéristiques $\mathbf{o}$ utilisées pour apprendre le CAC de phonétisation, des n-grammes de phonèmes sont aussi utilisés ici. Après l'apprentissage, la probabilité d'élision $\varepsilon(p, \mathbf{g}, \mathbf{o})$ d'un phonème donné p est obtenu à partir du CAC d'élision de la manière suivante :

$$\varepsilon(p, \mathbf{g}, \mathbf{o}) = \begin{cases} 0.5 \times \Pr(e|p, \mathbf{g}, \mathbf{o}) & \text{si } e = \textit{optionnel}, \\ 1 - \Pr(e|p, \mathbf{g}, \mathbf{o}) & \text{si } e = \textit{obligatoire}\ , \end{cases} \tag{1}$$

où e représente l'étiquette obtenue par le CAC d'élision pour p et $\Pr(e|p, \mathbf{g}, \mathbf{o})$ est sa probabilité *a posteriori*. Comme seulement deux étiquettes sont possibles et celle retournée est la plus probable, $\Pr(e|p, \mathbf{g}, \mathbf{o})$ est toujours dans l'intervalle $[0, 5; 1]$. D'après (1), $\varepsilon(p, \mathbf{g}, \mathbf{o})$ varie ainsi dans $[0; 0, 5]$. Cette définition permet ainsi d'éviter au modèle d'élision de complètement supprimer les choix faits par le modèle de phonétisation. Rien n'empêche néanmoins de l'adapter, par ex. pour y intégrer de connaissances *a priori*.

En conséquence, la probabilité d'un phonème peut être reformulée de la manière suivante :

$$\Pr(p|\mathbf{g}, \mathbf{o}) = \phi(p|\mathbf{g}, \mathbf{o}) \times (1 - \varepsilon(p, \mathbf{g}, \mathbf{o}))\ , \tag{2}$$

et la probabilité complémentaire d'élider p est :

$$\Pr(\epsilon|p, \mathbf{g}, \mathbf{o}) = \phi(p|\mathbf{g}, \mathbf{o}) \times \varepsilon(p, \mathbf{g}, \mathbf{o})\ , \tag{3}$$

où ϵ signifie l'absence de phonème. En utilisant ces probabilités, un treillis de phonèmes peut être créé pour chaque séquence de phonèmes donnée et le chemin avec la plus grande probabilité est choisi comme la meilleure prononciation. L'architecture d'un tel treillis est illusté par la figure 1a. Les arcs sont étiquetés par un phonème ou ϵ et sont alors respectivement pondérés par les probabilités de (2) ou (3). Ce principe peut être étendu aux meilleures hypothèses retournées par le modèle de phonétisation. Après application du modèle d'élision sur chaque hypothèse, un nouveau treillis peut être construit comme l'union de toutes les séquences alternatives de phonèmes.

3.3 Expériences sur des mots isolés

La méthode de conversion proposée a été appliquée sur le corpus MHATLex (Pérennou & De Calmes, 2000). Ce corpus comprend $450\,000$ mots avec un total de $710\,000$ prononciations. Chaque mot

possède une étiquette grammaticale et chaque prononciation inclut des possibilités d'élision ainsi que les contextes phonologiques pour lesquels chaque prononciation s'applique. Ce corpus est une version plus détaillée du corpus BDLex, utilisé dans (Illina *et al.*, 2011). Les contextes phonologiques sont ignorés pour la première série d'expériences. Ils seront pris en compte dans la section 4. Le corpus a été découpé en trois parties : ensembles d'apprentissage (75 %), de développement (5%), et de test (20%). Les 2 000 mots les plus fréquents du français ont été placés dans l'ensemble d'apprentissage car ces mots ne seront jamais des mots hors-vocabulaire dans des applications réelles, et possèdent de plus des prononciations irrégulières. De plus, les mots issus d'un même lemme ont été regroupés dans le même ensemble afin d'éviter aux différents ensembles d'être morphologiquement trop similaires. Les modèles CAC de phonétisation et d'élision sont appris sur l'ensemble d'apprentissage en utilisant l'ensemble de développement pour définir le critère d'arrêt tandis que les évaluations sont conduites sur l'ensemble de test. Wapiti est utilisé pour entraîner les CAC. Les graphèmes et phonèmes ont été alignés en utilisant un outil d'alignement plusieurs-à-plusieurs [2] et les CAC entraînés grâce à l'outil Wapiti [3] (Lavergne *et al.*, 2010).

Différents jeux de descripteurs ont été testés pour l'apprentissage du CAC de phonétisation. Ceux-ci rassemblent des n-grammes de graphème (pour rappel, un graphème g_i entouré par une fenêtre de $\pm W$ graphèmes) et l'information verbe/non verbe. Différentes tailles de fenêtre W ont été testées, tandis que l'information sur le verbe a toujours été utilisée. En plus de ces descripteurs, le modèle d'élision prend en compte le phonème p_i et ses $\pm W$ phonèmes voisins. W est fixé à la même valeur pour les graphèmes et les phonèmes afin d'éviter l'ajout d'un paramètre supplémentaire.

La table 1 présente les taux d'erreurs au niveau phonème (PER) et au niveau mot (WER) sur l'ensemble de test pour différentes tailles de fenêtre et avec ou sans modèle d'élision. Les résultats sont comparés à ceux obtenus par Liaphon, le système le plus utilisé pour la phonétisation d'énoncés pour le français (Béchet, 2001). Liaphon repose sur des règles manuelles qui couvrent les règles générales de prononciation ainsi que les exceptions. La version utilisée pour les expériences est une version modifiée optimisée pour l'usage en synthèse de parole. Au contraire, les résultats pour l'approche à base de CAC de (Illina *et al.*, 2011) ne sont pas reportés ici car le corpus et la stratégie de partitionnement des données sont différents. Premièrement, il apparaît que, pour $W = 2$, notre approche obtient des résultats proches de ceux de Liaphon, bien que légèrement moins bons. L'accroissement de la taille de la fenêtre des graphèmes apporte un gain. Cependant, après une taille de 2, il est apparu dans nos expériences que la qualité des CAC était dégradée. Cela vient probablement du fait que l'ensemble d'apprentissage contient beaucoup de mots assez proches en raison des contraintes sur les lemmes, ce qui amène à un effet de surapprentissage. Deuxièmement, l'usage d'un modèle d'élision amène de la variabilité dans le treillis de phonèmes sans significativement altérer ou améliorer les résultats.

4 Phonétisation d'énoncés

Dans cette section, nous proposons (i) de modéliser les transitions entre mots en introduisant la notion de contexte phonologique dans le cadre probabiliste posé précédemment et (ii) de calculer l'ensemble des variantes de prononciation d'un énoncé par la composition de TFP. Nous présentons la formalisation de ces contributions, puis leur validation expérimentale sur un corpus de parole.

2. `https://code.google.com/p/m2m-aligner/`
3. `https://wapiti.limsi.fr`

Descripteurs	PER (%)	WER (%)
Graphème (sans fenêtre) + verbe/non verbe	5,8	29,9
+ modèle d'élision	5,7	29,5
Graphème (± 1) + verbe/non verbe	2,6	11,3
+ modèle d'élision	2,4	11,6
Graphème (± 2) + verbe/non verbe	1,8	9,0
+ modèle d'élision	1,9	9,3
Liaphon	1,3	6,8

TABLE 1: PER et WER sur l'ensemble de test de MHATLex.

4.1 Introduction de contextes phonologiques

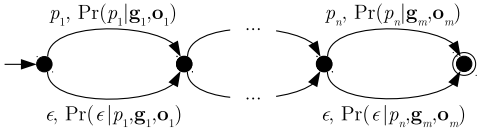
Un mot w_i influence la prononciation des mots précédent et suivant w_{i-1} et w_{i+1}. Réciproquement, la prononciation du mot w_i dépend de celle des mots w_{i-1} et w_{i+1}. L'influence des mots voisins est désignée ici comme le contexte phonologique. Soit l_i l'information transmise par w_i vers la gauche, c.-à-d. à w_{i-1}, et r_i l'information transmise vers la droite à w_{i+1}. De manière symétrique, la prononciation de w_i dépend de r_{i-1} et l_{i+1}. Ainsi, nous proposons d'intégrer r_{i-1} et l_{i+1} comme nouveaux descripteurs dans le processus d'apprentissage des CAC de phonétisation et d'élision.

4.2 Représentation sous forme de transducteurs finis pondérés

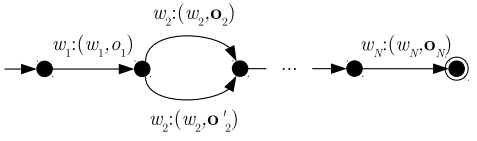
Pour calculer l'ensemble des variantes de prononciation d'un énoncé, nous proposons de construire un treillis de phonèmes en composant deux transducteurs finis pondérés : le premier représentant l'énoncé, le second toutes les prononciations possibles de ses mots.

Comme le montre la figure 1b, la représentation TFP d'un énoncé de N mots consiste simplement en un chaînage de nœuds dont les transitions transposent successivement chaque mot w_i en sa paramétrisation $(w_i, \mathbf{o}_i)$. Ce formalisme accepte d'éventuelles multiples paramétrisations pour un même mot, comme illustré avec le mot w_2 où des chemins alternatifs sont construits dans le transducteur. Par défaut, toutes les transitions ont une probabilité de 1.

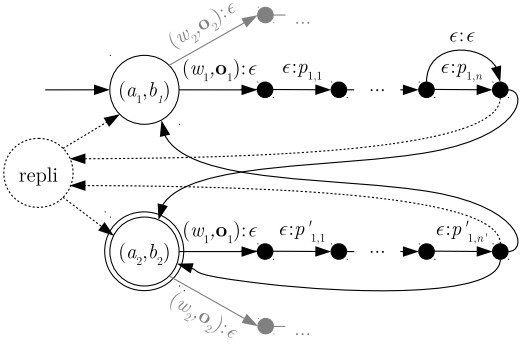
Le TFP du lexique de prononciation est plus complexe car les transitions entre mots doivent être modélisées. La figure 1c illustre son architecture. Pour chaque mot paramétré $(w_i, \mathbf{o}_i)$, plusieurs phonétisations peuvent être acceptables selon le contexte phonologique d'usage du mot. Ces contextes phonologiques sont représentés comme des nœuds (a, b) à partir desquels et vers lesquels chaque phonétisation est reliée. Entre ces nœuds, de même qu'à la figure 1a, chaque séquence de phonèmes est représentée comme une chaîne où $(w_i, \mathbf{o}_i)$ est consommé par le premier arc et les phonèmes $p_{i,j}$ sont les sorties des arcs restants. Les élisions sont représentées par des ϵ-transitions. Finalement, les transitions entre mots sont traitées de la manière suivante : chaque prononciation contextualisée $(r_{i-1}, w_i, \mathbf{o}_i, p_{i,1}, \ldots, p_{i,n}, l_{i+1})$ est liée à tous les nœuds contexte possibles (a, r_{i-1}) et (l_{i+1}, b), pour tous a et b de l'ensemble des contextes respectifs l_i et r_i transmis par w_i à gauche et à droite. Toutes les prononciations sont également liées à un nœud de repli pour autoriser des transitions théoriquement interdites. Les arcs vers les nœuds contexte sont pondérés avec une probabilité de 1 tandis ceux vers le nœud de repli sont pondérés avec une pénalité empirique fixée à e^{-10}. Enfin, en fonction de leur contexte phonologique, certains nœuds contexte sont définis comme terminaux. Les prononciations de chaque mot de l'énoncé sont soit dérivées à partir du dictionnaire, soit, pour les mots hors-vocabulaire, à partir du phonétiseur de mots en contexte présenté à la section 4.1. Pour



(a) Treillis de phonèmes pour un mot isolé.



(b) TFP des mots d'un énoncé.



(c) TFP d'un lexique de prononciation. Les arcs sans étiquette sont des transitions ϵ:ϵ. Les probabilités ne sont pas affichées par souci de clarté.

FIGURE 1

les prononciations dans le vocabulaire, la probabilité de chaque phonème est de 1 s'il est obligatoire et de $0,5$ s'il est optionnel. Dans ce second cas, une ϵ-transition de probabilité $0,5$ est construite en complément parallèlement au phonème. Les probabilités des mots hors-vocabulaire sont, elles, données par le phonétiseur de mots en contexte.

En composant le transducteur de l'énoncé avec celui du lexique, un treillis de phonèmes est obtenu et décodé pour générer la meilleure ou les meilleures prononciations pour l'énoncé.

4.3 Expériences sur des énoncés

Les valeurs possibles des contextes phonologiques sont issues du corpus MHATLex. Deux valeurs sont considérés pour r_{i-1} : l'une indique que le mot précédent se termine par une syllabe ouverte, l'autre par une syllable fermée. Celles pour l_{i+1} sont plus variées : le mot suivant peut débuter par une consonne, une semi-voyelle ou voyelle, un phonème nasal ou non, les liaisons peuvent être interdites ou alors il peut ne pas y avoir de mots suivant (fin de phrase). Ce dernier contexte est le seul cas permettant de définir un nœud contexte comme terminal. Les CAC de phonétisation et d'élision ont été réappris sur l'ensemble d'apprentissage de MHATLex augmenté des informations de contexte.

L'approche proposée a été appliquée sur un corpus de parole, dont la phonétisation a été vérifiée manuellement, d'environ $1\,400$ énoncés pour un total de $12\,000$ mots. Les énoncés ont été phonétisés avec la meilleure configuration de la section 3. Les résultats pour les quatre configuration testées sont mesurés en termes de PER et de taux d'erreurs sur les énoncés (SER, pour *Sentence Error Rate*).

Les résultats sont présentés dans la table 2 et sont comparés à ceux de Liaphon sur le même corpus. Tout d'abord, les PER sont bien plus élevés que sur les mots isolés. Cela montre clairement la difficulté de modéliser la prononciation d'énoncés. Ensuite, nous observons sur les différentes configurations que le modèle d'élision et les contextes phonologiques apportent des améliorations significatives qui, par ailleurs, se complètent en partie. Enfin, notre approche produit de moins bons résultats que Liaphon. Nous pensons que c'est logique car les treillis de phonèmes recensent de nombreux chemins équiprobables du fait de notre stratégie de pondération des prononciations issues du vocabulaire. Ces chemins sont notamment engendrés par des possibilités d'élision ou de liaison. En réalité, cette

Descripteurs et modèles	PER (%)	SER (%)
Graphèmes (± 2) + verbe/non verbe	$22,6$	$88,4$
+ modèle d'élision (sans contextes phonologiques)	$16,8$	$89,2$
+ contextes phonologique (sans modèle d'élision)	$17,7$	$85,6$
+ modèle d'élision + contextes phonologiques	$16,4$	$87,7$
Liaphon	$13,2$	$57,4$

TABLE 2: PER et SER sur le corpus de parole.

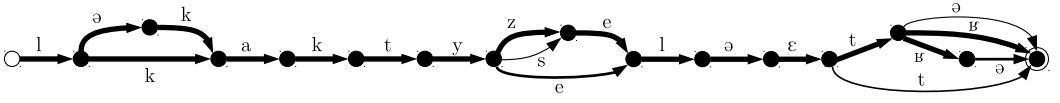


FIGURE 2: Treillis de phonèmes pour l'énoncé « le cactus et le hêtre ». Les transitions les plus probables sont représentées par les lignes les plus épaisses.

hypothèse d'équiprobabilité n'est pas vraie dans la langue. Au contraire, Liaphon fait des hypothèses concernant ces phénomènes. Il serait néanmoins simple de corriger cet effet dans notre méthode.

Un exemple de treillis de phonèmes (élagué) est donné par la figure 2 pour l'énoncé « le cactus et le hêtre », où les mots *cactus* et *hêtre* sont des mots hors-vocabulaire. Des chemins alternatifs apparaissent clairement. De meilleurs résultats pourraient sûrement être obtenus par différentes sophistications que nous prévoyons d'étudier, comme la prise en compte d'informations morphosyntaxiques plus riches (qui permettraient ici de corriger la phonétisation de *cactus*) ou une réévaluation du treillis de phonèmes en post-traitement par un modèle de langage. Néanmoins, cet exemple montre le potentiel de l'approche proposée.

5 Conclusions et perspectives

Cet article présente une nouvelle méthode de phonétisation du français. L'objectif principal de cette approche est de produire des treillis de phonèmes qui peuvent être facilement adaptés à des cas spécifiques, comme un style de parole ou un accent spécifique, en particulier pour la synthèse de parole. Cette méthode repose sur l'utilisation de champs aléatoires conditionnels pour phonétiser les mots isolés, élider certains phonèmes et prendre en compte les contextes phonologiques, ainsi que sur des transducteurs finis pondérés pour étendre la phonétisation à des énoncés.

De nombreuses perspectives sont offertes par cette approche. Tout d'abord, les TFP d'énoncés pourraient intégrer de multiples tokénisations ou prendre en compte des incertitudes sur la paramétrisation, par ex. au niveau des classes grammaticales. Cela peut notamment être utile pour des abréviations ou des acronymes. Ensuite, l'adaptation des treillis de phonèmes peut permettre d'améliorer les applications de la synthèse de parole où une expressivité ou un style de parole particuliers sont nécessaires, par ex. les jeux vidéo, les livres audio ou l'apprentissage de la langue. Enfin, l'utilisation des treillis par un moteur de synthèse de parole offrirait à ce dernier plus de choix et de flexibilité.

Références

BÉCHET F. (2001). LIA_PHON : un système complet de phonétisation de textes. *Traitement Automatique des Langues (TAL)*, (1).

BELLEGARDA J. R. (2005). Unsupervised, language-independent grapheme-to-phoneme conversion by latent analogy. *Speech Communication*, (2).

BENESTY J., SONDHI M. M. & HUANG Y. (2008). *Handbook of speech processing*. Springer.

BISANI M. & NEY H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*.

BODENSTAB N. & FANTY M. (2007). Multi-pass pronunciation adaptation. In *Proc. of ICASSP*.

CLAVEAU V. (2009). Letter-to-phoneme conversion by inference of rewriting rules. In *Proc. of Interspeech*.

HAHN S., LEHNEN P. & NEY H. (2011). Powerful extensions to CRFs for grapheme to phoneme conversion. In *Proc. of ICASSP*.

HAHN S., VOZILA P. & BISANI M. (2012). Comparison of grapheme-to-phoneme methods on large pronunciation dictionaries and LVCSR tasks. In *Proc. of Interspeech*.

HAZEN T. J., HETHERINGTON I. L., SHU H. & LIVESCU K. (2005). Pronunciation modeling using a finite-state transducer representation. *Speech Communication*, **46**(2).

ILLINA I., FOHR D. & JOUVET D. (2011). Grapheme-to-Phoneme Conversion using Conditional Random Fields. In *Proc. of Interspeech*.

JIAMPOJAMARN S., KONDRAK G. & SHERIF T. (2007). Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In *Proc. of HLT-NAACL*.

LAURENT A., DELÉGLISE P. & MEIGNIER S. (2009). Grapheme to phoneme conversion using an SMT system. In *Proc. of Interspeech*.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proc. of ACL*.

LEHNEN P., HAHN S., GUTA V.-A. & NEY H. (2012). Hidden conditional random fields with M-to-N alignments for grapheme-to-phoneme conversion. In *Proc. of Interspeech*.

LEHNEN P., HAHN S. & NEY H. (2011). N-grams for conditional random fields or a failure-transition (φ) posterior for acyclic FSTs. In *Proc. of Interspeech*.

MOHRI M., PEREIRA F. & RILEY M. (2000). Weighted finite-state transducers in speech recognition. In *Proc. of the Intl Workshop on Automatic Speech Recognition : Challenges for the Next Millenium*.

NOVAK J. R., MINEMATSU N. & HIROSE K. (2012). WFST-based grapheme-to-phoneme conversion : open source tools for alignment, model-building and decoding. In *Proc. of the 10th International Workshop on Finite State Methods and Natural Language Processing*.

PÉRENNOU G. & DE CALMES M. (2000). MHATLex : Lexical resources for modelling the French pronunciation. In *Proc. of LREC*.

POLYÁKOVA T. & BONAFONTE A. (2011). Introducing nativization to spanish TTS systems. *Speech Communication*, (8).

RAO K., PENG F., SAK H. & BEAUFAYS F. (2015). Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In *Proc. of ICASSP*.

WANG D. & KING S. (2011). Letter-to-sound pronunciation prediction using conditional random fields. *IEEE Signal Processing Letters*, (2).

YAO K. & ZWEIG G. (2015). Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. *ArXiv Computer Science - Computation and Language*.

Pics mélodiques prétoniques en portugais brésilien : une étude quantitative

Plínio A. Barbosa[1], Philippe Boula de Mareüil[2]
(1) Universidade Estadual de Campinas (Unicamp), Campinas, Brésil
(2) LIMSI, CNRS & Université Paris-Saclay, Orsay, France
pabarbosa.unicampbr@gmail.com, Philippe.Boula.de.Mareuil@limsi.fr

RESUME

Le présent travail porte sur un trait prosodique assez typique du portugais brésilien : un pic mélodique en position prétonique en fin d'énoncé déclaratif. Il vise à quantifier le phénomène, à partir d'enregistrements de cinq hommes et cinq femmes de l'état de São Paulo, en lecture et en narration. Il en résulte que des montées sur les prétoniques de 4 demi-tons suivies de descentes de 8 demi-tons, en moyenne, s'observent dans les deux styles de parole, chez les femmes. Chez les hommes, ces valeurs sont respectivement de 3 et 7 demi-tons. Ces montées-descentes d'une tierce et d'une quinte, respectivement, peuvent donner au portugais brésilien cette musicalité particulière et, puisque les descentes sont plus rapides chez les femmes, elles ouvrent des perspectives sociolinguistiques intéressantes.

ABSTRACT

Pre-stress pitch peaks in Brazilian Portuguese: a quantitative study

The reported work addresses a fairly typical prosodic feature of Brazilian Portuguese: a pre-stress pitch peak at the end of declarative utterances. This phenomenon is here quantified from the recording of read and narrated speech of five male and five female speakers from São Paulo State. Results show a great majority of pretonic melodic peaks, with on average rises of 4 (females) and 3 (males) semitones followed by pitch falls of 8 (females) and 7 (males) semitones, respectively, in both speaking styles. These rising/falling contours of about a third/fifth, respectively, may be responsible for the particular musicality of Brazilian Portuguese and, since pitch falls are faster in female speakers, they opens up an interesting avenue of investigation for sociolinguistics.

MOTS-CLES : prosodie, phonétique du portugais brésilien, clichés mélodiques
KEYWORDS: prosody, Brazilian Portuguese phonetics, melodic clichés

1 Introduction

Certaines langues, certains accents, comme par exemple le portugais brésilien, sont décrits comme chantants. Cette impression, quand elle ne relève pas purement d'un mythe, est difficile à quantifier, pour plusieurs raisons. Elle peut naître chez certains locuteurs et pas chez d'autres, dans certaines situations, dans certaines variétés de la langue et pas dans d'autres. Elle peut être diffuse ou ponctuelle, apparaissant uniquement en certains points de l'énoncé. Dans ce dernier cas, le contour nucléaire, le plus souvent autour du dernier accent lexical de la phrase — ou la clausule (Carton *et al*., 1991) — est un domaine approprié pour porter ce que Fónagy (1983) a appelé des *clichés* mélodiques. Ces clichés sont des patrons mélodiques de grande régularité qui créent des liens « directs et constants » entre une séquence verbale et une situation particulière (Fónagy *et al*., 1983). Une telle régularité pourrait concourir au caractère chantant de certaines variétés de langue.

Un exemple de cliché mélodique nous est offert par le portugais brésilien (dorénavant PB), avec un mouvement montant-descendant perceptivement très saillant, assez immédiatement audible dans différentes villes du Brésil et volontiers parodié par des humoristes français. Ce phénomène, peut-être surtout présent chez les femmes, a été traité dans plusieurs publications, et se montre relativement régulier, comme nous allons le voir.

Barone (2013) a rapporté des montées mélodiques saillantes suivies de descentes sur les syllabes toniques nucléaires, dans des phrases déclaratives, dans la variété de PB parlée à Recife (représentée par 3 hommes et 4 femmes). L'auteur a mesuré ce patron à la fin de phrases de structure sujet-verbe-objet (SVO). Dans le cas d'objets complexes (le plus souvent deux mots reliés par une préposition, comme dans *Rio de Janeiro*), en particulier, des pics anticipés ('early peaks') ont été relevés sur le dernier mot, chez les femmes notamment (79 % chez les femmes, 37 % chez les hommes). Ce patron a également été observé à Rio de Janeiro et à São Paulo (Frota & Moraes, 2016), sans référence à d'éventuelles différences liées au genre. Sur la base de données recueillies auprès d'une locutrice de Rio de Janeiro, Moraes (2008) propose de représenter la descente entre la syllabe prétonique et la finale tonique par un accent H+L*, suivi d'un ton de frontière L%, c'est-à-dire un niveau bas ('Low') ancré avec la syllabe tonique — notation également utilisée par Barone (2013). Dans toutes ces études, cependant, on ne trouve aucune mesure de combien sont les montées et les descentes, aucun commentaire sur la forme du contour nucléaire, aucune comparaison entre styles de parole, puisqu'elles reposent uniquement sur la lecture de phrases isolées, d'un petit nombre (ou un nombre non précisé) de locuteurs. Notre propre étude s'appuie sur des phrases lues connectées (d'un texte) et de la parole spontanée (ici restreinte à de la narration), de davantage de locuteurs et de locutrices. Nous limiterons le travail rapporté ici, toutefois, à la variété de PB de l'état de São Paulo.

La section suivante décrit le corpus utilisé et la méthode adoptée pour notre approche quantitative. La section 3 présente les résultats obtenus, comparativement pour les hommes et les femmes, dans deux situations de communication distinctes. Enfin, la section 4 conclut et ouvre quelques perspectives.

2 Corpus et méthode

Un texte d'environ 1600 mots, « O monge desastrado » 'Le moine désastreux', sur l'origine des pâtisseries nommées *pastéis* de Belém, a été lu par 10 locuteurs brésiliens (5 hommes, 5 femmes), suivi de la narration immédiate de l'histoire. On a ainsi constitué un corpus parallèle de lecture oralisée et de narration à partir de ce texte, originellement écrit en portugais européen et adapté au portugais brésilien (de plus traduit en français par les auteurs de cet article, dans le but de mener des études comparatives ultérieures). Le texte est particulièrement bien adapté pour fournir des contours intonatifs terminaux, dans la mesure où les phrases sont assez courtes. Afin d'examiner les contours censés être terminaux — de fin d'énoncé —, nous avons appliqué un critère simple : nous avons retenu les phrases terminées par un point ou un point d'exclamation dans le cas de la lecture. Après exclusion de quelques cas (tels des monosyllabes précédés d'une virgule), nous avons ainsi pu repérer 113 contours de fin de phrase, pour ce texte que chaque locuteur a lu. Les mots placés à la fin des phrases que nous avons retenues, pour l'analyse des clichés mélodiques, sont répartis come suit : 19 % sont oxytons, 77 % sont paroxytons et 4 % sont proparoxytons, c'est-à-dire que l'accent porte sur la dernière, l'avant-dernière ou l'antépénultième syllabe, respectivement. Cette répartition est proche de celle qui résulte de l'étude de Cintra (1997), sur le PB, dont les chiffres sont 70 % de paroxytons, 20 % d'oxytons et 10 % de proparoxytons.

En parole spontanée, il est moins évident de déterminer ce qui est un contour terminal : en narration, en particulier, on doit faire face à beaucoup d'énoncés continuatifs, où la voix reste en suspens, même à la fin de phrases pourtant syntaxiquement bien formées. Les auteurs (l'un natif du PB, l'autre parlant une variété plus proche du portugais européen) ont écouté conjointement la narration des 10 locuteurs et annoté ce qu'ils percevaient comme des frontières terminales, à la fin d'énoncés pragmatiquement complets et prosodiquement autonomes. Pour définir ces énoncés, ils se sont référés aux critères établis par Raso et Mello (2012), suivant la proposition de Cresti (2000) d'associer la propriété de terminalité à un acte illocutoire donné. Pour valider l'accord entre les auteurs, l'un des contributeurs à ces critères (T. Raso) a annoté indépendamment la narration la plus longue (près de 10 minutes). Ceci a permis de vérifier que les accords entre lui et les auteurs étaient presque parfaits, les rares exceptions venant de doutes à propos de contours mélodiques qui finissaient sur un ton élevé et étaient considérés par cet expert comme terminaux. Étant donné que ces quelques désaccords ne peuvent pas compromettre les tendances générales, nous avons poursuivi l'annotation sans aide extérieure.

Certains locuteurs se montrant peu bavards, d'autres produisant essentiellement des contours continuatifs, nous avons pu mesurer bien moins de contours terminaux en narration qu'en lecture. La table 1 résume la durée des textes lus et narrés, ainsi que le nombre de contours terminaux analysés, pour les hommes et les femmes.

La table 1 consigne en outre l'intervalle de temps moyen entre deux contours terminaux. Celui-ci est très variable en narration, allant d'environ 9 à 34 s, ce qui est évidemment lié à la façon de narrer de chaque locuteur ou locutrice. La locutrice AG, en particulier, a raconté l'histoire dans les moindres détails : elle le fait en environ cinq fois plus de temps que le deuxième locuteur le plus prolixe. Les énoncés lus, quant à eux, sont définis par le texte écrit, ce qui confère une régularité aux contours terminaux qu'illustrent les valeurs des intervalles moyens dans le tableau.

Sujet.Sexe	Lecture			Narration		
	durée (s)	#contours	Intervalle moyen (s)	durée (s)	# contours	Intervalle moyen (s)
AG.F	743,0	113	6,6	544,9	33	16,5
DF.F	656,2	113	5,8	99,2	4	24,8
GR.F	522,3	113	4,6	78,7	9	8,7
NP.F	599,0	113	5,3	84,5	6	14,1
RA.F	679,0	113	6,0	26,9	2	13,5
CA.M	761,1	113	6,7	68,6	8	8,6
EM.M	573,1	113	5,1	64,0	4	16,0
FA.M	635,0	113	5,6	68,8	5	13,8
LC.M	428,7	113	3,8	102,9	3	34,3
MT.M	509,3	113	4,5	125,7	6	21,0

TABLE 1 : Durée, nombre de contours et intervalle de temps moyen entre deux contours terminaux dans les extraits analysés (F=femmes, M=hommes).

Les contours terminaux que nous avons examinés ont été segmentés en syllabes — plus précisément, en unités allant de l'attaque d'une voyelle à l'attaque de la voyelle subséquente. Diverses études ont montré la pertinence de ces unités (également plus facile à délimiter) pour des études sur le rythme (Dogil & Braun, 1988 ; Barbosa, 2006 ; Pettorino *et al*., 2013, *inter alia*). Cette segmentation a été faite semi-automatiquement à l'aide du logiciel Praat (Boersma & Weenink, 2015), qui nous a également servi à extraire la fréquence fondamentale (F_0). Le script *BeatExtractor* (Barbosa, 2006) a dans un premier temps permis de repérer les débuts de voyelles ; après quoi, moyennant un très petit nombre de corrections manuelles, des symboles phonétiques ont été attribués manuellement à la chaîne de parole. Les maxima et minima de F_0 ont de même été annotés sous Praat, avec l'ancrage temporel des pics et vallées mélodiques dans, avant ou après la voyelle prétonique (ou une autre voyelle). La figure 1 en donne un exemple, en guise d'illustration.

À partir de ces annotations, les moyennes des montées (p-b) et des descentes (a-p) ont été calculées, des histogrammes des contours réalisés par les locuteurs (par pas d'1 demi-ton) ont été dressés, et les proportions de mouvements excédant un seuil donné ont pu être établies. Nous avons enfin regardé plus en détail si tel ou tel patron accentuel (oxyton, paroxyton ou proparoxyton) ou les frontières de mots favorisaient ou défavorisaient un pic sur la prétonique.

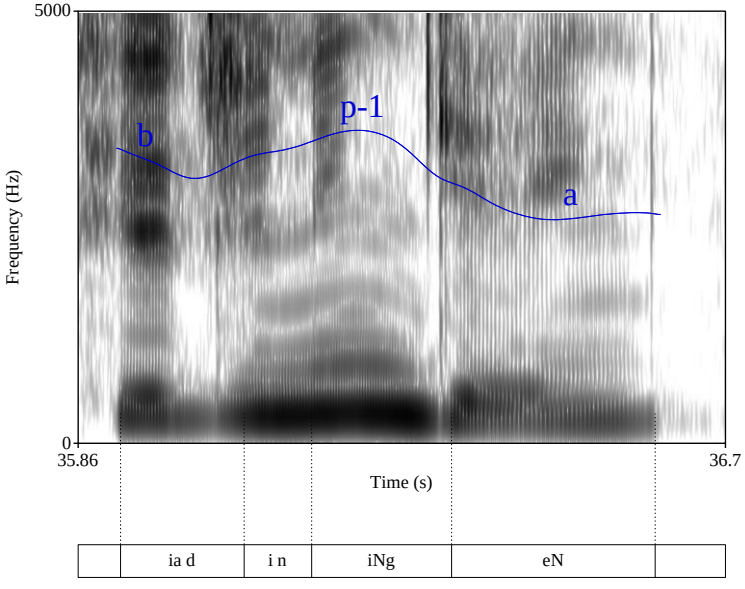


FIGURE 1 : Extrait du texte lu par la locutrice AG, pour le contour terminal associé à *de ninguém* 'de personne'. Le symbole b correspond au minimum de F_0 avant le pic mélodique ; p-1, marquant la position du pic, indique que celui-ci est ancré dans la syllabe précédant la tonique ; enfin le symbole a correspond au minimum de F_0 après le pic mélodique.

3 Résultats

Il est à noter que, chez tous les locuteurs, les pics mélodiques se trouvent à plus de 90 % en syllabe précédant immédiatement la tonique (position p-1), dans les deux styles de parole et ceci indépendamment du patron accentuel des mots. La table 2 rapporte, pour chaque locuteur, les moyennes et écarts types des montées et descentes ainsi que le pourcentage de montées prétoniques supérieures à 3 demi-tons. D'après 't Hart (1981), en effet, ce seuil est considéré comme une bonne estimation des corrélats acoustiques de proéminences prosodiques.

On voit dans cette table que les femmes varient moins que les hommes en lecture, tant pour les montées que pour les descentes. En valeurs moyennes, dans les deux styles de parole, les femmes montent et descendent d'un demi-ton de plus que les hommes. Ces valeurs moyennes se rapprochent entre hommes et femmes, pour les descentes en narration. Tant en lecture qu'en narration, les femmes montrent un pourcentage de montées supérieures à 3 demi-tons qui est d'environ le double de celui des hommes, atteignant plus de 80 % chez certaines locutrices.

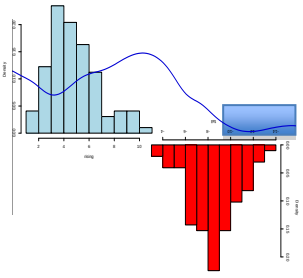
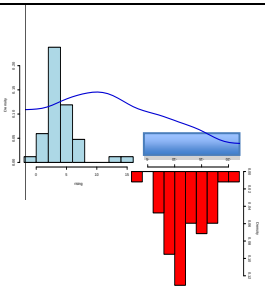
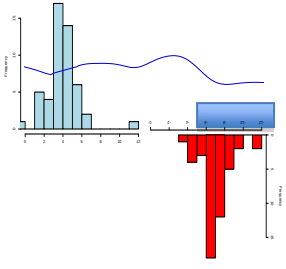
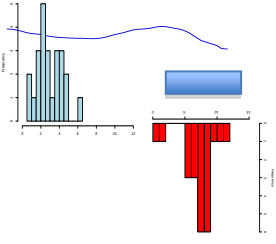
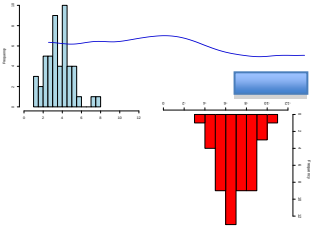
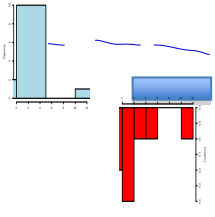
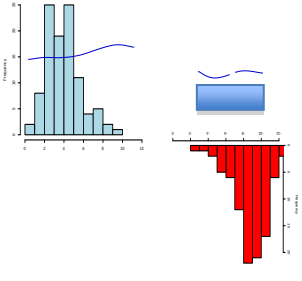
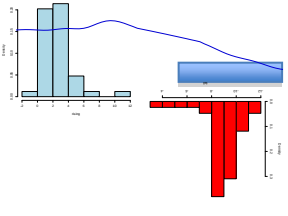
Sujet.Sexe	Lecture			Narration		
	montées	descentes	%montée > 3 dt	montées	descentes	%montée > 3 dt
AG.F	4,8 (2,0)	-8,4 (2,0)	84	6,0 (2,4)	-11,0 (3,0)	94
DF.F	3,9 (1,7)	-8,0 (2,2)	80	3,8 (3,0)	-5,5 (3,9)	50
GR.F	3,7 (1,4)	-6,9 (1,5)	69	4,3 (2,0)	-6,1 (1,9)	67
NP.F	4,0 (1,8)	-9,0 (2,0)	66	2,6 (1,0)	-5,3 (1,8)	60
RA.F	4,7 (1,6)	-9,3 (2,6)	89	4,5 (0,9)	-8,0 (0,9)	100
Moyenne femmes	**4,2 (1,7)**	**-8,3 (2,1)**	**78**	**4,2 (1,9)**	**-7,2 (2,3)**	**74**
CA.M	4,0 (2,9)	-12 (3,7)	67	3,6 (1,2)	-9,3 (4,7)	63
EM.M	3,0 (1,4)	-7,3 (2,5)	42	2,3 (0,8)	-3,6 (1,8)	25
FA.M	1,9 (7,0)	-0,1 (7,6)	40	3,3 (2,7)	-8,4 (5,7)	50
LC.M	2,5 (1,9)	-8,7 (1,6)	26	4,1 (3,1)	-6,1 (1,3)	33
MT.M	2,1 (0,7)	-7,4 (2,9)	10	2,2 (2,0)	-5,4 (1,1)	25
Moyenne hommes	**2,8 (2,8)**	**-7,1 (3,5)**	**37**	**3,1 (2,0)**	**-6,6 (2,9)**	**39**

TABLE 2 : Moyennes (et écarts types) des montées et des descentes en demi-tons (dt), et pourcentages de montées supérieures à 3 demi-tons (F=femmes, M=hommes).

La figure 2 présente les histogrammes des montées et descentes en demi-tons, ainsi que les formes typiques des contours pour tous les locuteurs, en lecture. Ces formes typiques ne sont pas des moyennes mais des contours réels (parmi les quelque 113 disponibles) dont les montées et les descentes correspondent aux valeurs moyennes pour un locuteur donné. Le rectangle à droite du contour représente l'extension et la position de la syllabe accentuée de ce même contour. Le but de ces diagrammes est non seulement de montrer la distribution des montées et descentes mais aussi de pouvoir comparer visuellement la rapidité des descentes chez les femmes et chez les hommes.

On observe que la descente est plus rapide chez les femmes que chez les hommes. De fait, la mesure de cette vélocité, en lecture, donne une chute moyenne de 34,0 dt/s chez les femmes contre 31,7 dt/s chez les hommes.

On note par ailleurs que les pics prétoniques affectent les mots qu'ils soient oxytons ou paroxytons (les proparoxytons étant très peu nombreux dans notre corpus) et qu'ils ne sont pas bloqués par une frontière de mot précédant immédiatement un accent lexical. Généralement, les montées et descentes se font à l'intérieur d'un syntagme formé de deux mots, dont le premier peut être : un clitique, le cas le plus courant (ex. *de ferro,* 'en fer'*, os lábios, '*les lèvres'*, a ele,* 'à lui'), un adverbe (ex. *bem longe,* 'bien loin') ou un déterminant indéfini (ex. *nenhuma falta, '*aucune faute'). On peut tout à fait observer des pics sur des clitiques précédant des paroxytons dissyllabiques, comme dans les exemples cités ci-dessus. Cette configuration est très fréquente, alors que dans une langue comme le français, par exemple, une montée mélodique sur le clitique est très rare (Boula de Mareüil *et al*., 2011).

	AG		CA
	DF		EM
	GR		FA
	NP		LC

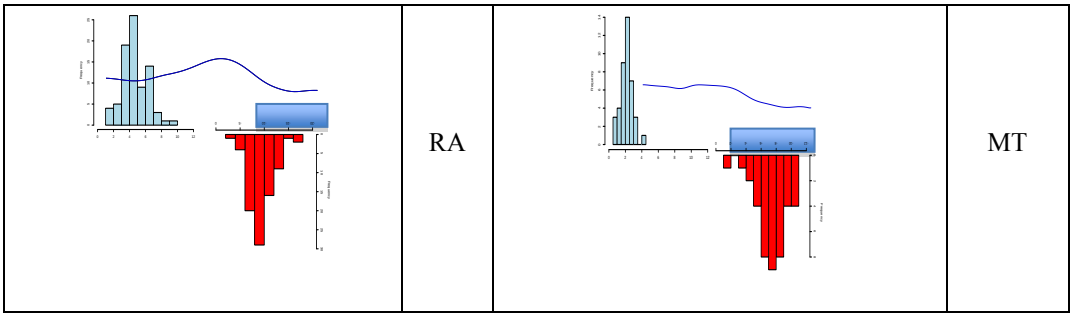


FIGURE 2 : Histogrammes des montées (en bleu) et descentes (en rouge), avec les contours typiques, en lecture, de tous les locuteurs (à droite) et toutes les locutrices (à gauche).

4 Conclusion et perspectives

Cet article, centré autour de la mesure, en production, de contours intonatifs terminaux en portugais brésilien, a mis en évidence une tendance à faire précéder le dernier accent de l'énoncé par une montée mélodique (de 3–4 demi-tons), qui se prolonge par une descente (de 7–8 demi-tons), de façon relativement régulière et récurrente, en lecture comme en parole spontanée — malgré une moindre représentativité de cette dernière. Si l'on fait référence à des échelles musicales, ces intervalles correspondent respectivement à une tierce et à une quinte, ce qui n'est sans doute pas étranger à l'impression de langue chantante que peut donner le portugais brésilien. Le taux plus élevé de montées supérieures à 3 demi-tons et des descentes plus rapides, chez les femmes, contribuent très certainement à renforcer cette impression de la part des locutrices. Des expériences perceptives sont nécessaires pour étayer cette perception d'intervalles musicaux transposés dans le domaine de la parole. Elles pourraient s'inspirer de Ferreira Netto *et al*. (2013), dont le travail considère des seuils de perception. Elles devraient également inclure davantage de locuteurs (hommes et femmes), d'autres points d'enquête au Brésil.

D'après nombre d'études sociolinguistiques, les femmes font figure de pionnières en matière d'innovations linguistiques (Labov, 2001). Le fait que le contour montant-descendant que nous avons mesuré sur la prétonique apparaisse comme un trait féminin plus que masculin demande donc à être étudié de plus près. Le même corpus, déjà enregistré auprès de locuteurs de portugais européen, devraient dans un avenir proche nous renseigner sur la spécificité brésilienne du patron prosodique étudié. Celui-ci ne s'observe pas au Portugal, d'après nos premières écoutes. Enfin, le même protocole ayant été appliqué en français, langue à maints égards différente sur le plan prosodique, d'autres comparaisons seront possibles.

Remerciements

Le travail présenté ici a été financé par le programme « Chaires franco-brésiliennes dans l'état de São Paulo », un partenariat entre l'Unicamp et le Consulat Général de France à São Paulo accordé au second auteur. Le premier auteur développe sa recherche grâce à l'octroi d'une bourse brésilienne du CNPq #301387/2011-7. Les auteurs sont tous deux reconnaissants envers Robson R. Monteiro et Tommaso Raso pour leur aide précieuse à cette étude.

Références

BARBOSA, P. A. (2006). *Incursões em torno do ritmo da fala*. Pontes, Campinas.

BARONE, M. (2013). A comparative study on high pre-tonic pitch accents between a Brazilian Portuguese and an Italian variety: a case of supra-segmental reanalysis, *IV Colóquio Brasileiro de Prosódia da Fala*, Maceió, 1–5.

BOERSMA, P. et WEENINK, D. (2015). Praat: doing phonetics by computer (Version 5.4.22) [Logiciel]. Téléchargé de : <http://www.praat.org/>

BOULA DE MAREÜIL, P., RILLIARD, A., ALLAUZEN, A. (2011). A diachronic study of initial stress and other prosodic features in the French news announcer style: corpus-based measurements and perceptual experiments. *Language and Speech* 55(2), 263–293.

CARTON, F., ESPESSER, R., VAISSIERE, J. (1991). Étude sur la perception de l'"accent" régional du Nord et de l'Est de la France. *12e Congrès International des Sciences Phonétiques*, Aix-en-Provence, 422–425.

CINTRA, G. (1997). Distribuição de padrões acentuais no vocábulo em português. *Confluência* 5(3), 82–93.

CRESTI, E. (2000). *Corpus di italiano parlato*, vol. I-II, [CD-ROM]. Accademia della Crusca, Florence.

DOGIL, G. et BRAUN, G. (1988). *The PIVOT model of speech parsing*, Academie Verlag, Vienne.

FERREIRA NETTO, W., PERES, D. O., MARTINS, M.V.M., ROSA, R.C.M., VIEIRA, M. F. (2013). Análise automática de manifestações emocionais de tristeza e cólera em PB: abordagem pelo programa ExProsodia. *Leitura* 52, 43–65.

FÓNAGY, I. (1983). *La vive voix. Essais de psycho-phonétique,* Payot, Paris.

FÓNAGY, I., BÉRARD, E., FÓNAGY, J. (1983). Clichés mélodiques. *Folia linguistica* 17, 153–185.

FROTA, S. et MORAES, J. A. de (2016). Intonation of European and Brazilian Portuguese. *In* Wetzels, W. L., Menuzzi, S., Costa, J. (éditeurs), *The Handbook of Portuguese Linguistics*. Malden: Willey-Blackwell, sous presse.

LABOV, W. (2001). *Principles of linguistic change. Social factors*, Blackwell, Oxford.

MORAES, J. A. de (2008). The pitch accents in Brazilian Portuguese: Analysis by synthesis. *4th International Conference on Speech Prosody*, Campinas, 389–397.

PETTORINO, M., MAFFIA, M., PELLEGRINO, E., VITALE, M. DE MEO, A. (2013). VtoV: A perceptual cue for rhythm identification. *International Prosody-Discourse Interface Conference*, Leuven, 101–106.

RASO, T., MELLO, H. [Org.] (2012). *C-ORAL-BRASIL I. Corpus de referência do português brasileiro falado informal*, Editora UFMG, Belo Horizonte.

'T HART, J. (1981). Differential sensitivity to pitch distance, particularly in speech. *Journal of the Acoustical Society of America* 69(3), 811–821.

Préservation du pattern syllabique iambique dans la production des locuteurs dysarthriques

Laurianne Georgeton [1] Christine Meunier [2]
(1) Police Technique et Scientifique, Ministère l'Intérieur, France
(2) Laboratoire Parole et Langage, Aix Marseille Université, CNRS
5 avenue Pasteur, 13100 Aix-en-Provence, France
laurianne.georgeton@gmail.com, christine.meunier@lpl-aix.fr

RESUME

Ce travail vise à évaluer une éventuelle dégradation du pattern rythmique iambique dans la production de locuteurs atteints de différents types de dysarthrie. Ce pattern se traduit par une structure court-long dans les mots dissyllabiques. Cette structure est très robuste en français aussi bien en production qu'en perception. Par ailleurs, chez des locuteurs dysarthriques, des perturbations prosodiques et donc rythmiques sont souvent observées. Ainsi, ces patients peuvent-ils maintenir ce pattern iambique dans leurs productions? Les résultats montrent que le pattern rythmique iambique est bien conservé chez toutes les populations dysarthriques aussi bien en lecture qu'en parole spontanée. Ce pattern est en général plus marqué en spontané qu'en lecture et la population contrôle se démarque des populations dysarthriques par un pattern plus marqué en lecture, mais plus encore en spontané. Ce pattern rythmique semble donc robuste même s'il semble être affecté quand la sévérité de la maladie augmente.

ABSTRACT

The preservation of iambic syllabic pattern in the production of dysarthric speakers.
This study aims to evaluate a potential degradation of the iambic rhythmic pattern in the production of speakers who suffer from various dysarthric pathologies. This pattern is characterized by a short-long structure within dissyllabic words. It's strongly robust in French both in production and perception. Furthermore, prosodic and rhythmic distortions are commonly observed in the productions of dysarthric speakers. Thus, are they able to maintain the iambic pattern in their production? Results show that the iambic pattern in properly preserved for all the dysarthric populations both in read and spontaneous productions. The pattern effect is stronger in spontaneous than in read speech. Moreover, controlled speakers show a more pronounced pattern, especially in spontaneous speech. Thus, this rhythmic pattern seems to have robust components, even if distortions appear when pathology reaches high degree of severity.

MOTS-CLES: Dysarthrie, rythme, parole spontané, lecture.
KEYWORDS: Dysarthria, rhythm, spontaneous speech, reading task

1 Introduction

Ce travail vise à évaluer une éventuelle dégradation des patterns rythmiques fondamentaux au niveau de la syllabe en français dans la production de locuteurs atteints de différents types de

dysarthrie. De nombreux travaux ont pu mettre en évidence le rôle important des aspects temporels dans l'intelligibilité et la compréhension de la parole (Fraisse, 1956). Ce rôle tient à une organisation prosodique complexe portant sur de larges segments linguistiques, mais aussi à une structure rythmique concernant la production des mots et des *Accentual Phrase* (AP, Jun & Fougeron, 2002). Cette structuration rythmique occasionne un allongement caractéristique des syllabes finales des AP et donc très souvent des syllabes finales des mots. Ainsi, la production de mots pluri-syllabiques se caractérise systématiquement par un allongement de la syllabe finale de ces mots. Cette information est traitée par les locuteurs pour la segmentation en mots lors de la perception. Banel et Bacri (1993) ont ainsi pu montrer que lors d'écoute de séquences dissyllabiques ambigües (« corps beau » vs « corbeau »), les auditeurs identifient deux mots lorsque le pattern rythmique proposé est de type trochaïque (long-court) et un seul mot lorsqu'il est de type iambique (court-long). Il semble donc que ce pattern rythmique soit robuste en français aussi bien en production qu'en perception. Il faut également remarquer que ce pattern est préservé en parole spontanée conversationnelle alors que ce type de parole présente de fortes différences sur le plan rythmique par rapport à la parole lue. Dans plusieurs études (Adda-Decker et al., 2008 ; Meunier et Espesser, 2011), un allongement final marqué sur les mots pluri-syllabiques en parole conversationnelle a pu être mis en évidence. Ces résultats suggèrent que le pattern iambique est très robuste pour ce qui concerne la structure rythmique des mots dissyllabiques en français.

L'objectif de ce travail est donc d'observer si ce pattern robuste est, ou non, préservé chez des populations atteintes de pathologies affectant le contrôle moteur et donc l'activité motrice liée à la production de la parole. La dysarthrie résulte d'une atteinte du système nerveux central ou périphérique affectant la réalisation motrice de la parole. Ainsi, suivant les pathologies, la parole est caractérisée par une hypoarticulation ayant pour conséquence une production imprécise des consonnes (Kent et al., 1991) et une réduction articulatoire des voyelles pour certaines populations (Audibert & Fougeron, 2012). La parole est également affectée au niveau de la prosodie (mélodie et rythme) : elle peut être très ralentie (SLA : Weismer et al., 2000) ou au contraire rapide et marquée par de nombreuses pauses (Parkinson : Skoda and Schlegel, 2008). L'ataxie cérébelleuse se caractérise par un défaut de coordination des gestes et une accentuation fluctuante (Gilman & Klein, 1992), ce qui entraîne une parole ralentie mais également irrégulière aussi bien dans l'articulation que dans la temporalité. L'ensemble de ces observations nous conduit à rechercher, dans les productions pathologiques, les paramètres phonétiques qui sont préservés et ceux qui ne le sont pas. En effet, connaître les paramètres qui sont préservés le plus tard possible dans les pathologies de la parole nous fournit une information sur les caractéristiques fondamentales et robustes d'une langue. On peut supposer que la plupart des locuteurs tendent à maintenir le plus longtemps possible (de façon stratégique et adaptative) les caractéristiques de la langue qui leur permettent le plus efficacement d'être compris par leur entourage. Ainsi, dans ce travail, nous proposons d'évaluer la préservation du pattern syllabique iambique chez trois populations de locuteurs atteints de dysarthrie (maladie de Parkinson, ataxie cérébelleuse et sclérose latérale amyotrophique).

La tâche de production (i.e. lecture d'un texte ou production de parole spontanée) peut induire des différences dans les réalisations des locuteurs dysarthriques. Pour certains la tâche de lecture semble facilitatrice tandis que pour d'autres elle semble engendrer une contrainte. Ces différences se manifestent par le fait que certains patients sont jugés avec un degré de sévérité plus important en lecture qu'en spontané alors que c'est l'inverse pour d'autres (Lhoussaine, 2012). Pour cette raison, et pour observer éventuellement le rôle du style de parole sur la préservation du pattern iambique, ce travail porte sur l'analyse de deux types de parole distincts : la lecture d'un texte et la parole spontanée de type narrative.

2 Méthode

2.1 Corpus et populations

Trois populations dysarthriques et un groupe de locuteurs sains ont été comparés, soit quatre groupes différents : 12 locuteurs atteints de Sclérose Latérale (SLA), 8 patients atteints de la maladie de Parkinson (PARK), 8 patients atteints d'ataxie cérébelleuse (ATAX) et 6 locuteurs sains (CTRL). La moyenne d'âge est comparable entre les groupes.

Populations	Age	Degrés de Sévérité	
		Lecture	Spontanée
CTRL	69 <63-82>	x	x
PARK	64.5 <48-83>	0.84 <0.37-1.37>	0.99 <0.36-1.64>
SLA	66 <50-81>	2.05 <0.91-2.91>	2.02 <1.18-2.73>
ATAX	55 <32-77>	1.28 <0.82-2.09>	1.23 <0.64-1.9>

TABLE 1 : Ages et degrés de sévérité en lecture et en spontanée pour chaque population étudiée. CTRL: sujets contrôles, PARK: sujets atteints de Parkinson, SLA: sujets atteints de SLA, ATAX: sujets atteints d'ataxie cérébelleuse. Moyenne <min-max>

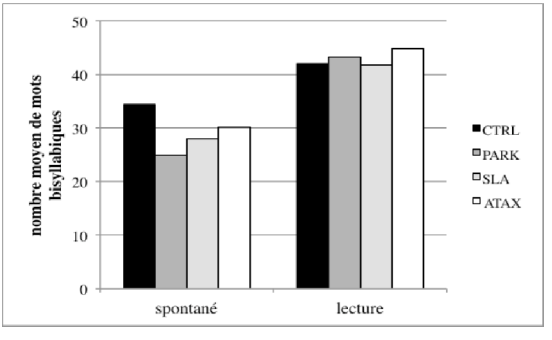


FIGURE 1: Nombre moyen de mots dissyllabiques produits en tâche spontanée et en tâche de lecture des populations CTRL, PARK, SLA, ATAX.

Les enregistrements ont été extraits de bases de données élaborées dans le cadre de deux projets ANR[1]. Deux situations de parole ont été considérées : 1/ la lecture d'un texte de 168 mots (« le cordonnier ») et 2/ la parole en conversation dans une situation d'interview (récits d'événements du quotidien). La production en interview est de durée très inégale : assez longue pour les locuteurs CTRL (8 min en moyenne, parole totale sans pause), plutôt courte pour les locuteurs dysarthriques (50 secondes en moyenne, parole totale sans pause). La plupart des locuteurs dysarthriques évitent les situations de productions de parole dans laquelle ils se sentent mal à l'aise. Pour pouvoir comparer les populations entre elles, nous avons fait le choix de ne prendre en

[1] ANR DesPhoAPady, Description Phonético-Acoustique de la Parole Dysarthrique, Projet ANR-08-BLAN-0125.
ANR TYPALOC, Variations normales et anormales de la parole: TYPologie, Adaptation, LOCalisation, Projet ANR-12-BSH2-003.

compte que les 50 premières secondes de parole chez les sujets CTRL. Le nombre moyen de mots dissyllabiques produits sont globalement comparables entre les populations, comme le montre la Figure 1.

2.2 Mesures

Les corpus enregistrés ont été transcrits, puis alignés automatiquement et enfin corrigés par un expert. La syllabation du corpus a été réalisée grâce au logiciel SPPAS (Bigi, 2015, http://www.sppas.org/) reposant sur trois principes : 1/ une syllabe contient une voyelle et seulement une ; 2/ une pause est une frontière de syllabe et 3/ une frontière de mot est une frontière de syllabe. Par exemple, la séquence [dãzɛ̃pəti] est segmentée [dãz .ɛ̃.pə.ti] et non [dã.zɛ̃.pə.ti].

La durée des premières et deuxièmes syllabes (S1 et S2) a été extraite de l'ensemble des mots dissyllabiques dans les corpus des deux situations de parole pour les quatre populations. Pour quantifier la différence entre S1 et S2, nous avons calculé le ratio suivant : S2/S1.

2.3 Analyses statistiques

Les analyses statistiques ont été menées sur le logiciel R en utilisant des modèles à effets mixtes (package lme4, Pinheira & Bates, 2000). Les analyses statistiques ont été menées séparément selon les tâches de production i.e. tâche en spontanée vs tâche de lecture. Nous avons testé un modèle intégrant les facteurs fixes suivants : « Population » (4 niveaux : CTRL, PARK, SLA, ATAX) et « rang de la syllabe » (2 rangs : syllabe 1 (S1) et syllabe 2 (S2)) et leur interaction avec la durée des syllabes. Les résultats montrent une interaction significative entre le rang de la syllabe et la population en tâche de lecture (χ^2 (7)= 447, p=.001) et en tâche de production spontanée (χ^2 (7)= 530, p=.001). Des comparaisons multiples entre les deux facteurs « Population » et « Rang de la syllabe » ont été menées avec la fonction *glht* du package *multcomp*. Le seuil de significativité, déterminé à $p<.05$ a été obtenu par la méthode d'approximation de Satterthwaite. Les résultats post-hoc sont présentés dans la partie 3.1. Nous avons également testé l'effet de la « Population » (facteur fixe) sur le ratio S2/S1 (variable dépendante) avec le facteur « locuteurs ». Les tests statistiques montrent qu'il existe un effet de la population sur le ratio en production spontanée (χ^2 (3)= 8,5, p=.004*) et en tâche de lecture (χ^2 (3)= 10.2, p=.002*). Les résultats post-hoc sont présentés dans la partie 3.2.

3 Résultats

Les résultats sont structurés en trois parties. Les valeurs et différences temporelles entre S1 et S2 sont commentées dans un premier temps pour les quatre populations et dans les deux styles de parole. Le rapport entre S2 et S1 est ensuite analysé de façon à évaluer l'importance du pattern pour les quatre populations et dans les deux styles de parole. Enfin, et pour affiner un peu les résultats, nous commentons les résultats de quelques locuteurs spécifiques.

3.1 Valeurs temporelles de S1 et S2

Les résultats montrent que, pour les mots dissyllabiques, la seconde syllabe est significativement plus longue que la première syllabe pour les sujets contrôles et pour chaque population de patients dysarthriques. Cependant, quelques différences sont notables entre les populations selon la tâche

de production. En tâche de production spontanée, la différence de durée entre la première et la seconde syllabe sont comparables entre les populations dysarthriques (i.e. log(0.5) msec), comme le montre la Table 3, alors que les locuteurs contrôles sont caractérisés par un allongement plus élevé de la seconde syllabe (i.e. log(0.7) msec). En tâche de lecture, la différence de durée entre les locuteurs atteints de Parkinson et les locuteurs contrôles est similaire (i.e. log(0.34) msec). Comparé à ces deux populations, l'allongement de la seconde syllabe est moins élevé pour la population ATAX (i.e. log(0.27) msec) et pour la population SLA (i.e. log(0.16) msec).

	pop	spontaneous		reading	
		S1	S2	S1	S2
Dissyllabes	CTRL	139 (78)	272 (132)	185 (62)	266 (108)
	PARK	143 (73)	222 (88)	179 (60)	251 (88)
	SLA	259 (182)	367 (185)	339 (170)	386 (162)
	ATAX	188 (96)	295 (114)	243 (86)	317 (105)

TABLE 2 : Valeurs moyennes et écart-types (entre parenthèses) des syllabes 1 et 2 des mots dissyllabiques en tâche de production spontanée et en tâche de lecture pour chaque population (CTRL, PARK, SLA, ATAX), en msec.

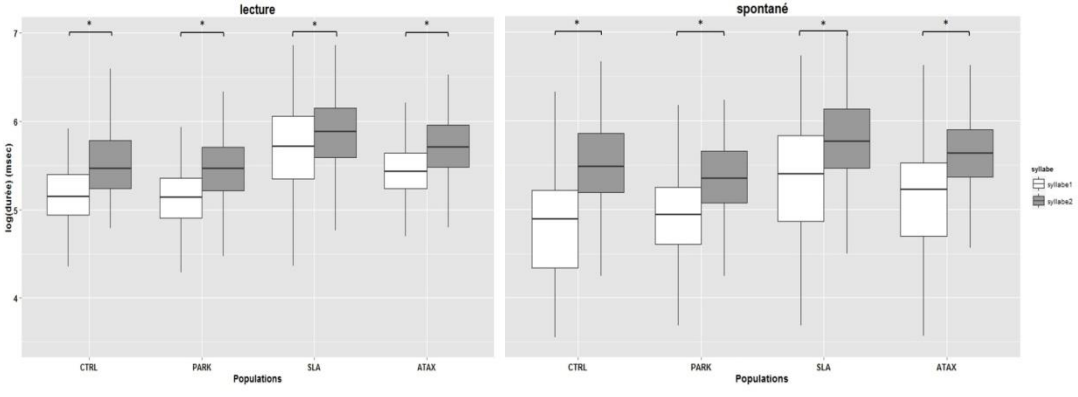


FIGURE 2: Durées de S1 (en blanc) et de S2 (en gris) des mots dissyllabiques en lecture (à gauche) et en production spontanée (à droite) pour 4 populations : population de référence (CTRL) et populations dysarthriques (PARK, SLA, ATAX), en log (durée) msec.

	Populations	Différence S2-S1 (z value, * :p<0.05) (log(durée) en msec)	
		Spontané	lecture
Dissyllabes	CTRL	0.72 ± 0.05 (14.5 *)	0.34 ± 0.03 (11.1 *)
	PARK	0.50 ± 0.05 (9.9 *)	0.34 ± 0.03 (12.9 *)
	SLA	0.48 ± 0.04 (12.4 *)	0.16 ± 0.02 (7.5 *)
	ATAX	0.51 ± 0.04 (11.2*)	0.27 ± 0.03 (10.4 *)

TABLE 3 : Valeurs moyennes et écart-types (entre parenthèses) des syllabes 1 et 2 des mots dissyllabiques en tâche de production spontanée et en tâche de lecture pour chaque population (CTRL, PARK, SLA, ATAX), en log(durée) msec.

Comme le montre la Figure 2, nous observons plus de variabilité dans la tâche de production spontanée que dans la tâche de lecture. Toutefois, cette variabilité n'est pas due à une variation de la structure syllabique de S1 ni de S2 dont la distribution est très homogène au travers des quatre populations. On notera aussi un contraste plus marqué en spontané qu'en lecture : l'écart entre S1 et S2 et bien plus important en parole spontanée et cela pour toutes les populations, aussi bien contrôle que dysarthriques. La production spontanée semble donc induire un relief rythmique plus contrasté dans lequel les populations dysarthriques parviennent à préserver le pattern iambique. En revanche, la tâche de lecture semble induire une production moins contrastée dans laquelle les locuteurs dysarthriques peinent plus à préserver le pattern iambique (voir plus bas).

3.2 Ratio S2/S1

Afin de quantifier ces différences entre la première et la seconde syllabe et de déterminer si les populations se distinguent les unes des autres, nous avons calculé un ratio S2/S1. En tâche de production spontanée, les tests post-hoc montre que chaque population dysarthrique se distingue significativement de la population contrôle (différence CTRL-PARK : -0.52 ± 0.2 (z : 2.7*), différence CTRL-CER : -0.6 ± 0.2 (z : 2.8*), différence CTRL-SLA : -0.52 ± 0.2 (z : 2.5*)).

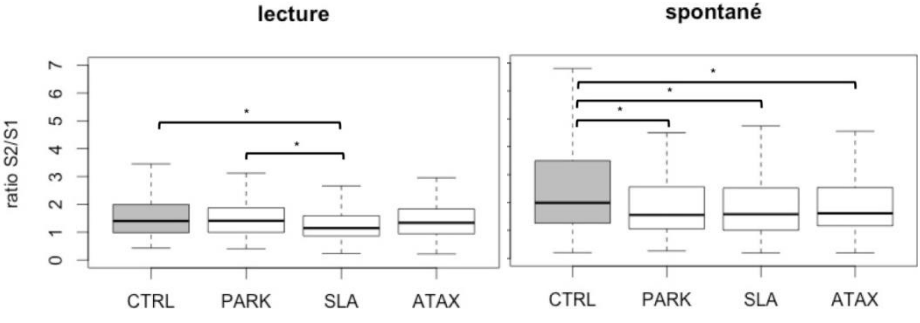


FIGURE 3: Ratio S2/S1 pour les mots dissyllabiques dans une tâche de lecture (à gauche) et dans une tâche de production spontanée (à droite) pour 4 populations : population de référence (CTRL) et populations dysarthriques (PARK, SLA, ATAX).

En tâche de lecture, la population SLA se distingue significativement de la population contrôle (différence CTRL-SLA : -0.3 ± 0.09 (z : 3*)) et de la population atteinte de la maladie de parkinson (différence PARK-SLA : -0.2 ± 0.08 (z : 2.6*)).

3.3 Patterns interindividuelles

Les résultats présentés ci-dessus représentent des valeurs globales regroupant l'ensemble des locuteurs de chaque population. Toutefois, les populations pathologiques sont très souvent caractérisées par une forte variabilité interindividuelle. Cette variabilité est due à de multiples facteurs parmi lesquels on peut évoquer les spécificités de l'atteinte physiopathologique, le profil psychologique (réaction face à la maladie), les stratégies de compensation qui peuvent être très différentes selon les ressources de chaque individu. Ainsi, si l'on observe les réalisations des locuteurs contrôles dans la tâche de lecture (figure 4), on constate une certaine homogénéité dans la production des mots dissyllabiques et dans le respect du pattern iambique.

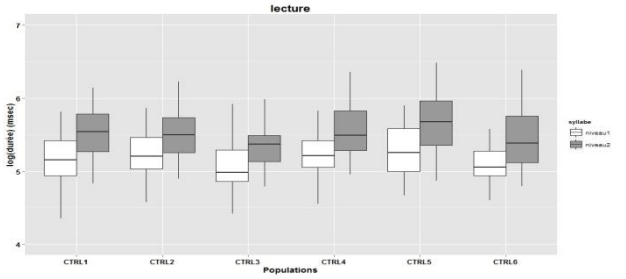


FIGURE 4: Durées de S1 (en blanc) et de S2 (en gris) des mots dissyllabiques dans la tâche de lecture pour les 6 locuteurs CTRL.

Il en va différemment pour la population des locuteurs SLA pour lesquels une forte variabilité interindividuelle peut être observée dans la tâche de lecture (figure 5). Nous avons choisi de présenter cette population car c'est celle qui présente le plus de contraste entre locuteurs. C'est aussi la population pour laquelle de degré de sévérité est le plus élevé. Plus spécifiquement, on constate que trois locuteurs ne parviennent pas à préserver le pattern rythmique iambique (figure 5, entouré de rouge). Ces trois locuteurs, SLA1, SLA4 et SLA8 sont également ceux qui montrent le degré de sévérité le plus élevé (respectivement : 2,9 ; 2,4 et 2,6 sur une échelle de 0 à 3). Il semble donc qu'au-delà d'une certaine gravité de la maladie, les patterns rythmiques fondamentaux n'arrivent plus à être préservés entraînant ainsi une perte d'intelligibilité considérable. Notons que ces trois locuteurs parviennent à préserver le pattern iambique en tâche de parole spontanée, ce qui suggère que la contrainte de la tâche de lecture ne favorise pas les contrastes rythmiques.

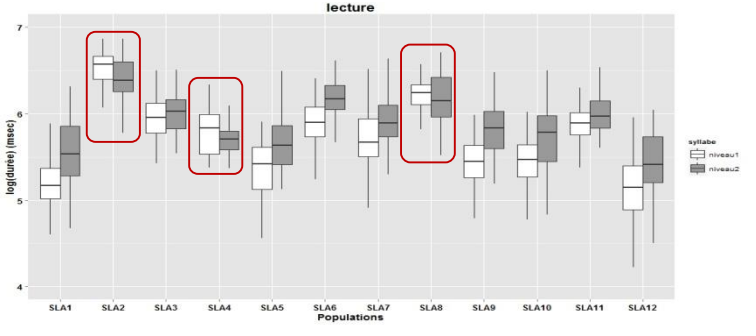


FIGURE 5: Durées de S1 (en blanc) et de S2 (en gris) des mots dissyllabiques dans la tâche de lecture pour les 12 locuteurs SLA.

4 Discussion

Cette étude visait à évaluer la capacité de locuteurs atteints de trois types de dysarthrie à maintenir le pattern rythmique iambique sous-jacent à la production des mots plurisyllabiques en français. Nos questions se basaient 1/ sur la robustesse du pattern iambique en français (Jun & Fougeron, 2002) ; 2/ la dégradation des structures prosodiques et rythmiques dans les productions des locuteurs dysarthriques. Il s'agissait donc de comprendre si la robustesse de ce pattern constituait

un socle suffisamment important pour la communication pour que les locuteurs dysarthriques tentent de le maintenir le plus longtemps possible. L'ensemble des résultats montre une présence nette et régulière du pattern iambique dans les mots dissyllabiques chez toutes les populations (contrôle et dysarthriques). Pour toutes ces populations la différence entre S1 et S2 est significative, avec une durée plus longue pour S2. On note que les débits spécifiques des populations (plutôt lent pour les SLA et rapide pour les Park) apparaissent en parallèle du pattern iambique, mais ne le perturbe pas. Le pattern est présent aussi bien en lecture qu'en parole spontanée. Toutefois, il apparaît plus marqué en parole spontanée (rapport S2/S1 plus important) et cela pour toutes les populations. Les écarts temporels et rythmiques sont clairement majorés en parole spontanée, tandis que la lecture de texte semble induire un nivellement des différences rythmiques. Ce nivellement est visible dans le rapport S2/S1 où l'on voit que la population contrôle se distingue moins nettement des autres populations (même si la différence reste significative) que dans le contexte spontané où le rapport S2/S1 est très important. Ainsi, la différence entre S1 et S2 pourrait devenir plus ténu en lecture et, en conséquence, plus délicate à maintenir pour les locuteurs dysarthriques lorsque le trouble moteur devient trop sévère. C'est effectivement ce que l'on observe chez les patients SLA pour lesquels le rapport S2/S1 diminue significativement par rapport aux autres populations. Notons que la population SLA est celle pour laquelle nous recensons les degrés de sévérité les plus forts. En analysant plus en détail les locuteurs SLA en lecture on constate une très forte variabilité inter-individuelle sur plusieurs aspects : premièrement pour ce qui concerne le débit, ensuite concernant la variabilité et enfin concernant le pattern iambique. Pour ce troisième aspect, on observe que certains locuteurs marquent peu la distinction entre S1 et S2. Mais plus encore, trois locuteurs montrent un pattern inverse avec une S1 plus longue que la S2. Ces deux locuteurs montrent des degrés de sévérité les plus forts de la population SLA. Il semble donc que le pattern iambique soit bien maintenu chez les populations dysarthriques mais que, lorsque la maladie devient trop sévère, le pattern pourrait être affecté, comme l'ensemble des autres paramètres de la production.

En résumé, nous avons vu que le pattern rythmique iambique est bien conservé chez toutes les populations dysarthriques aussi bien en lecture en parole spontané. Ce pattern est plus marqué en spontané qu'en lecture et la population contrôle se démarque des populations dysarthriques par un pattern plus marqué en lecture, mais encore plus en spontané. Ce pattern rythmique semble donc très robuste.

Les locuteurs d'une langue partagent une compétence de la langue (grammaire intérieure). Les travaux sur les pathologies de la parole nous permettent ainsi de mettre en évidence ce qui, dans cette compétence, se révèle fondamental pour la communication, ou au contraire accessoire. Des travaux ont en effet montré que des locuteurs dysarthriques omettent de façon plus importante des segments variables et instables (/r/) ou facultatifs (liaisons, Meunier, 2015). Il y aurait donc, pour chaque locuteur, y compris ceux qui souffrent de déficit moteur, un savoir intuitif de ce qu'il faut préserver et de ce qui est modulable.

Remerciements

Ce travail a pu être réalisé grâce au soutien financier du projet TYPALOC (ANR-12-BSH2-003).

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* 49 (3), 13–46.

AUDIBERT N., & FOUGERON C. (2012). Distorsions de l'espace vocalique : quelles mesures? Application à la dysarthrie. Actes des Conferences JEP-TALN-RECITAL, Grenoble, 217–224.

BANEL M.H., BACRI N. (1993). Reconnaissance de la parole et indices de segmentation métriques et phonotactiques. *L'année psychologique*, 97, 77-112.

BIGI B. (2015). SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech. *The Phonetician*, 111-112, 55-69.

FRAISSE P. (1956). *Les structures rythmiques: étude psychologique*. Publications universitaires de Louvain.

GILMAN S. & KLUIN K. J. (1992). Speech disorders in cerebellar degeneration studied with positron emission tomography. *Neurologic disorders of the larynx.* A. Blitzer, M. F. Brin, C. T. Sasaki, S. Fahn & K. S. Harris, editors, 279-285.

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

KENT R., SUFIT R., ROSENBEK J., KENT J. WEISMER G., MARTIN R. (1991). Speech deterioration in amyotrophic lateral sclerosis: a case study. *Journal of Speech and Hearing Research*, 34, 1269-1275.

LHOUSSAINE L. (2012). Première validation de la Grille dEvaluation Perceptive de la Dysarthrie (G.E.P.D.) : effet du niveau dexpertise du jury et diffrenciation entre types de dysarthrie. Ph.D. Dissertation. Speech therapist thesis, University of Paris VI, Pierre er Marie Curie.

MEUNIER C., DOLCEMASCOLO A., FAURE M., GEORGETON L. (2015). Localisation ciblée des réductions phonétiques dans la dysarthrie. Actes des 6*ème* *Journées de Phonétique Clinique*, Montpellier, France.

MEUNIER C., ESPESSER R. (2011). Vowel reduction in conversational speech in French: The role of lexical factors. *Journal of Phonetics*, 39 (3), 271-278.

PINHEIRO J.C. AND BATES D.M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag, New York.

SKODDA S., SCHLEGEL U. (2008). Speech rate and rhythm in Parkinson's disease. *Movement Disorders*, 23(7), 985–992.

WEISMER G., LAURES J., JENG J.Y., KENT R., KENT J. (2000). Effect of speaking rate manipulations on acoustic and perceptual aspects of the dysarthria in amyotrophic lateral sclerosis. *Folia phoniatrica et logopaedica*, 52(5), 201–219.

Production des voyelles parlées et chantées dans le *Cantu in Paghjella*

Claire Pillot-Loiseau[1] Patrick Chawah[1] Angélique Amelot[1] Grégoire Bachman[1] Catherine Herrgott[1] Martine Adda-Decker[1] Lise Crevier-Buchman[1,2]
(1) Laboratoire de Phonétique et Phonologie UMR 7018, Université Paris 3 Sorbonne Nouvelle, 19 rue des Bernardins, 75005 Paris, France
(2) Hôpital Européen G. Pompidou, Univ. Paris Descartes, 20 rue Leblanc 75015 Paris
claire.pillot@univ-paris3.fr, patrick.chawah@gmail.com, angelique.amelot@univ-paris3.fr, gregoire.bachman@univ-paris3.fr, catherine.herrgott@gmail.com, madda@univ-paris3.fr, lbuchman@numericable.fr

RESUME

Quelles sont les caractéristiques acoustiques et articulatoires des voyelles parlées et chantées du *Cantu in Paghjella* (polyphonie corse à trois voix), en fonction du chanteur, de la voyelle et de la fréquence fondamentale ? L'analyse acoustique des quatre premiers formants de la parole au chant et celle des mouvements articulatoires lingual et labial, montrent généralement (i) une significative augmentation de F1 avec abaissement lingual mais fermeture labiale, en lien avec une corrélation entre F0 et F1 ; (ii) une baisse de F2 pour les voyelles antérieures, une postériorisation linguale et un recul de l'ombre hyoïdienne uniquement pour le *bassu* ; (iii) une nette augmentation de F3 et F4 surtout chez le *bassu* ; (iv) une augmentation du *Singing Power* Ratio surtout chez les *bassu* et *secunda*. Ses valeurs sont toutefois inférieures à celles de chanteurs lyriques, et ne correspondant pas comme ces derniers à un rapprochement de F3 et F4.

ABSTRACT

Production of spoken and sung vowels in *Cantu in Paghjella*
What are the acoustic and articulatory characteristics of spoken and sung vowels in the *Cantu in Paghjella* (Corsican polyphony for three voices), depending on the singer, the vowel and the fundamental frequency? The acoustic analysis of the first four formants in speech and singing, and the tongue and lips movements analysis usually show (i) a significant increase in F1, with decrease of tongue height but decrease of labial opening, in connection with a correlation between F0 and F1; (ii) a decrease of F2 for the front vowels, a lingual posteriorisation and a posteriorisation of the hyoid shadow only for the *bassu*; (iii) an increase of F3 and F4 mostly for the *bassu*; (iv) an increase of the *Singing Power Ratio*, mainly for the *bassu* and the *secunda*. Its values are however lower than those of opera singers and do not correspond to a clustering of F3 and F4.

MOTS-CLES : *Cantu in Paghjella*, voyelles, parole, chant, formants, SPR,contour lingual, lèvres
KEYWORDS: *Cantu in Paghjella*, vowels, speech, singing, formants, SPR, lingual contour, lips

1 Introduction et état de l'art

La tradition corse comprend plusieurs types de chants polyphoniques : le *madrigale*, hérité d'Italie, sur des thèmes souvent amoureux, les *terzetti*, où le motif poétique domine, et la *paghjella*, la plus ancienne (Catinchi, 1999). Une *paghjella* est une strophe de vers de huit syllabes traditionnellement chantée par trois hommes, de trois tessitures différentes. Les trois voix entrent de manière quasi

immuable : débute d'abord *l'a secunda*, voix principale, ténor, chantant la mélodie, puis *l'u bassu*, voix de basse, qui soutient et accompagne *l'a secunda*, et enfin entre *l'a terza*, voix la plus aigue, apporte des ornementations appelées mélismes ou *ricuccate*. (Bithell, 2007).

Le corse est une langue romane de l'aire italique (groupe italo-roman). Ses voyelles sont, en position tonique : /i/, /e/, /ɛ/, /a/, /ɔ/, /o/ et /u/. Dans le quart nord-est de la Corse, émerge un huitième phonème, /æ/, issue essentiellement de variantes de /ɛ/ ou /a/ devant nasale ou r + consonne. En position atone, l'inventaire vocalique comprend trois voyelles (/i/, /a/ et /u/) dans le Sud de la Corse, quatre au nord (/i/, /ɛ/, /u/, et /a/) et cinq au centre-est de la Corse (/i/, /ɛ/, /a/, /ɔ/, et /u/). Le corse présente une nasalisation vocalique qui peut aboutir à la phonologisation de véritables voyelles nasales. Enfin, l'alternance vocalique existe par jeu de déplacement de l'accentuation qui entraîne la modification de la voyelle devenue prétonique (Dalbera-Stefanaggi, 2002).

Les modifications phonétiques vocaliques de la parole au chant ont principalement été étudiées pour le chant classique dans un contexte monodique au niveau acoustique (Sundberg 1987 ; Bloothooft et Plomp, 1984) ou articulatoire (Sundberg 2009) et pour des fréquences fondamentales (F0) aigues (Titze *et al.* 1994 ; Sundberg et Skoog 1997), mais aussi dans d'autres techniques vocales (entre autres : Burns 1986 ; Gibson 2010). Ces auteurs affirment qu'en chant lyrique, les voyelles chantées sont en moyenne produites plus ouvertes, avec F1 plus élevé, F2 plus bas pour les voyelles antérieures, et un rapprochement de F3 et F4 (descendant de la parole au chant) pour produire le formant du chanteur. D'un point de vue articulatoire, pour cette même technique vocale, on observe une ouverture mandibulaire croissante avec une fréquence fondamentale plus aigue, ainsi qu'une réduction de la hauteur linguale pour /i, e, u, a/ (Sundberg 2009 ; Nair *et al.* 2016).

Nous cherchons à savoir quelles sont les caractéristiques acoustiques et articulatoires des voyelles parlées et chantées du *Cantu in Paghjella* en fonction du sujet qui chante, de la voyelle, et de la fréquence fondamentale (F0). Cette technique vocale en plein air ou dans les lieux de fête, n'a, à notre connaissance, jamais fait l'objet d'une analyse phonétique de ses productions vocaliques, en contexte polyphonique. Nous chercherons à savoir si les valeurs des quatre premiers formants et celles du *Singing Power Ratio* (*cf.* plus loin) évoluent en fonction des facteurs cités ci-dessus. Nous présentons également quelques données articulatoires montrant la configuration linguale et labiale grâce à la plateforme multicapteurs développée dans le projet *iTreasures* (Chawah *et al.* 2014).

2 Méthode

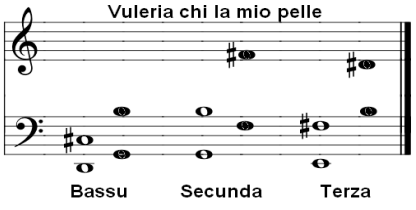
Deux groupes de trois chanteurs de *Paghjella* ont été enregistrés en Corse, à leur domicile : les chanteurs ont été répartis sur trois pièces différentes pour assurer une isolation acoustique nécessaire pour nos études. Pour qu'ils puissent s'écouter en chantant, des dispositifs radio-fréquentiels munis d'écouteurs ont été mis à leur disposition et la synchronisation était établie par un frappement de mains au début de l'enregistrement, mais ce mode d'enregistrement ne permettait pas aux chanteurs de se voir. Au dire d'un de ces sujets cependant, le groupe se connaît bien et a donc pu effectuer la performance demandée. Seuls le *secunda* et le *bassu* ont été enregistrés aux niveaux acoustique et articulatoire, car nous ne possédions que deux casques multicapteurs. Tous les sujets sont originaires du nord de la Corse, ou Haute-Corse. Les *secunda* et *terza* sont des corsophones maniant le corse quotidiennement et l'ayant appris par transmission orale dans leurs familles respectives en même temps que le français. Le *bassu*, plus jeune, manie moins la langue et a notamment appris la technique du chant polyphonique au cours de stages et ateliers.

Nous avons choisi de porter nos analyses directement sur les chants produits par nos sujets. Sur trois chants analysés, nous ne présentons ici que le chant *Vuleria chì la mio pelle.* Les paroles complètes de ce chant en trois couplets, seulement chantées en totalité par le *secunda*, étaient les suivantes (à

gauche : normal : *secunda* seul ; souligné : *secunda+bassu* ; italique : *secunda+terza* ; gras : voyelles choisies pour les données articulatoires ; majuscule : voyelles accentuées) :

« - VuleRIA CHI LA mio p***E****LLE* — *- Je voudrais que ma peau*
*DIventA*ssI Un *c**o**ghjU fOrte* — *Devienne un cuir fort*
- Per mandAlla A la c*Oncia* — *- Pour l'envoyer chez les tanneurs*
PEr fAnne Un *p**A**ghju dI bOtte* — *Et en faire une paire de bottes*
- Per pudE pUrtAlle tU — *- Pour que tu puisses les porter*
L*U mio aMO*re *fin'à la MOrte* » — *Mon amour, jusqu'à la mort.*

La figure 1 montre les étendues en fréquence de ce chant, complétée par les données quantitatives.



Voix (sujet)	Bassu		Secunda		Terza	
Tâche	**P**	**Cht**	**P**	**Cht**	**P**	**Cht**
Min (Hz)	72	100	98	174	80	241
Max (Hz)	140	247	246	369	180	316
Moyenne (Hz)	107	195	186	259	132	292
Médiane (Hz)	106	203	189	262	134	300
E-type (Hz)	13	42	30	42	20	19
Nombre	147	120	141	141	129	87

FIGURE 1: Etendues par voix du chant produit en voix parlée (rondes : P) et chantée (rondes noires : Cht) dans la même session : à gauche: données en notes ; à droite: données quantitatives en Hz avec le nombre d'occurrences vocaliques analysées pour chaque chant.

Les données acoustiques et physiologiques ont été acquises grâce à la plateforme multicapteurs développée dans le cadre du projet européen *iTreasures* (Chawah *et al.* 2014), avec un casque portable muni d'une sonde microconvexe 8MC4X (Vermon) reliée à un module Terason T3000, pour permettre l'acquisition des images ultrason linguales sur un ordinateur via un câble FireWire. Sur ce casque se trouvait également, devant les lèvres de chaque sujet, une caméra vidéo (modèle DFM 22BUC03-ML, CMOS USBmono) pour acquérir les images labiales. Enfin, un micro serre-tête cardioïde électrostatique (modèle C520L, marque AKG) était aussi utilisé pour enregistrer le signal acoustique. Un accéléromètre piézoélectrique (K&K Sound), un électroglottographe (EG2-PCX2 Glottal Enterprises) et une ceinture respiratoire ont également été utilisés au cours de ces enregistrements, mais non utilisés dans les analyses ici présentées. Un module nommé *i-THRec* (*i-Treasures Helmet Recording* software) a spécialement été conçu pour créer les calibrations adéquates et les sessions d'enregistrement. Il sert d'interface et est une surcouche du logiciel RTMaps® (Intempora inc.) assurant la synchronisation de l'acquisition des différents signaux et images. Les données ainsi obtenues pouvaient alors être directement traitées sur une interface graphique MATLAB nommée *i-THan* (*i-Treasures Helmet Analysis* software).

Au préalable a été annoté chaque vers du chant au niveau orthographique puis phonétique, via l'accord de deux annotatrices expertes en écoute phonétique, mais ne connaissant pas la langue corse. Les valeurs de fréquence fondamentale (F0) à 25%, 50% et 75% de chaque occurrence vocalique parlée et chantée, ainsi que les fréquences des quatre premiers formants aux mêmes points, ont été ensuite mesurées avec un script Praat (Boersma *et al.*, 2016) en permettant l'extraction automatique des valeurs, le réglage formantique étant comme seuil maximal de 5000Hz. Cette détection automatique a ensuite été vérifiée manuellement. La durée moyenne de chaque voyelle a également été détectée. Le *Singing Power Ratio* (Omori *et al.* 1996 : différence d'amplitude entre l'harmonique le plus élevé entre 2000 et 4000Hz, et l'harmonique le plus élevé entre 0 et 2000Hz) a également été mesuré. Cette valeur, négative, est inférieure pour la parole et les non chanteurs, par rapport aux chanteurs lyriques (Omori *et al.* 1996 ; Pillot et Vaissière, 2007).

Pour ce faire, chaque voyelle a été séparée de l'ensemble et un script Matlab a permis de détecter les valeurs formantiques (jusque F6), puis la différence a été effectuée sous Excel (moyenne sur chaque voyelle).

La visualisation qualitative des contours de la langue à 25%, 50% et 75% (ainsi que de l'ombre hyoïdienne) s'est effectuée grâce au script Mattong (Fux *et al.* 2014), celle des contours labiaux aux mêmes points avec un autre script Matlab sur les voyelles en gras des paroles du chant sélectionné ci-dessus. Les analyses statistiques ont utilisé des corrélations et des ANOVA avec le logiciel R.

3 Résultats

3.1 Valeurs formantiques

La figure 2 montre que F1 augmente significativement de la parole au chant et qu'il diffère d'un sujet à l'autre. F2 diminue pour les voyelles antérieures de parlé à chanté (mais non significativement pour /i/) et augmente légèrement pour les voyelles postérieures pour les *bassu* et *secunda*. Les valeurs de F3 et F4 sont très variables et augmentent chez le *bassu* quelle que soit la voyelle (significativement pour /a/). On n'observe pas de rapprochement entre F3 et F4 sauf pour le /a/ du *bassu*. Le tableau 1 indique que F0, F1, F2 et F3 de /a/, et F4 sont significativement différents en fonction de la voix (sujet) et de la tâche (parlé/chanté). Il existe également une interaction des facteurs voix x parlé/chanté pour la fréquence fondamentale (F0), F2 de /a/, F3 et F4 de /u/.

Mesures rép.	F0	F1			F2			F3			F4		
Voyelle		i	a	u	i	a	u	i	a	u	i	a	u
Voix	**451.5**	**18.1**	**28.5**	**7.2**	2.1	**5.9**	1.8	0.4	**40.2**	**10**	**6.5**	**7.5**	**5.6**
Parlé/chanté	**2172.2**	**127.8**	**368.2**	**67.5**	0.1	**34.4**	2.4	0.8	**55.9**	3.4	**16.6**	**17.7**	**7.2**
Voix x parlé/chanté	**136.2**	0.8	3.4	3.1	0.6	**27**	1.7	**9.4**	**17**	7.5	*4.8*	1.1	**8.8**

TABLE 1 : Valeur de F (arrondie au dixième) d'une ANOVA calculée pour les mesures répétées des facteurs voix (sujet), parlé/chanté, et voix x parlé chanté pour F0 et les valeurs des quatre premiers formants de /i/, /a/ et /u/. Gras : p< 0.00001 ; normal : non significatif.

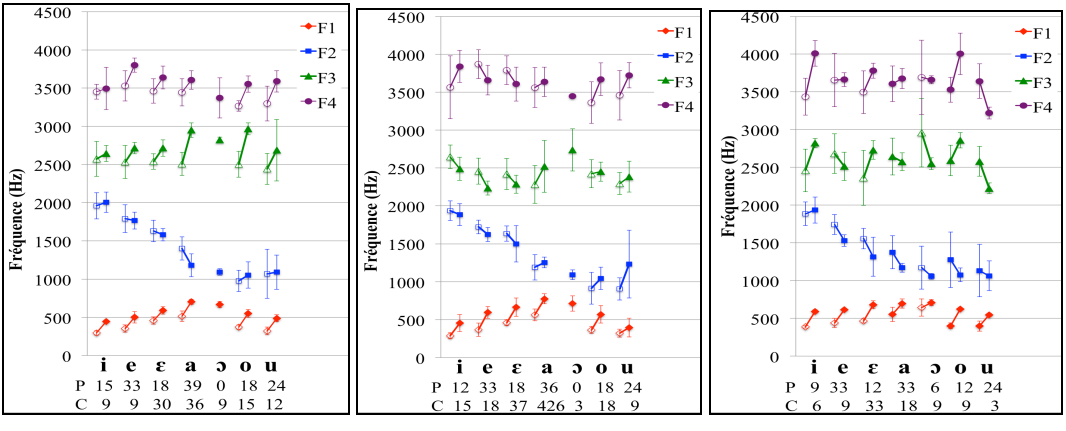


FIGURE 2: Moyennes, écart-types et nombre d'occurrences pour F1 à F4 (Hz) pour les *bassu* (gauche), *secunda* (milieu) et *terza* (droite) pour chaque voyelle parlée (forme vide, P) et chantée (forme pleine, C).

3.2 *Singing Power Ratio* (SPR) et Spectres Moyennés à Long Terme (LTAS)

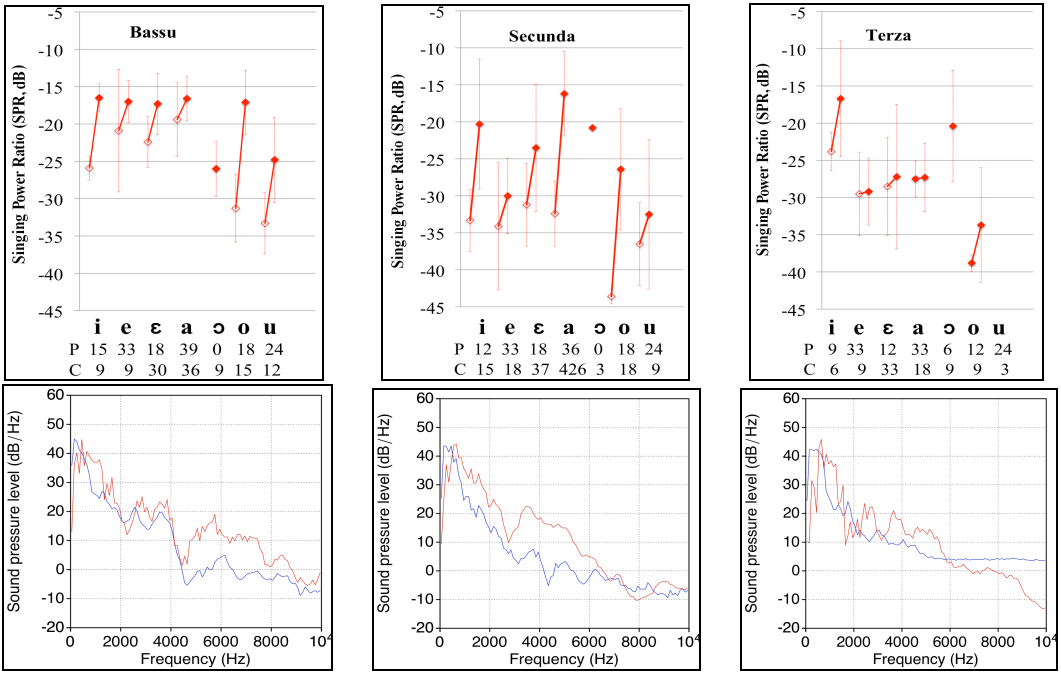


FIGURE 3: haut : moyennes, écart-types et nombre d'occurrences du SPR (dB) pour les *bassu* (gauche), *secunda* (milieu) et *terza* (droite) pour chaque voyelle parlée (forme vide, P) et chantée (forme pleine, C) ; bas : LTAS (tous segments voisés confondus) du chant mesuré pour les *bassu* (gauche), *secunda* (milieu) et *terza* (droite). Parlé : bleu ; chanté : rouge. Largeur de bande : 100Hz.

La figure 3 montre des valeurs très variables de SPR, qui toutefois augmente significativement de la parole au chant pour /i/ dans les trois voix ($F=22,3$; $p<0,0001$), /a/ chez le *secunda* ($F=18,9$; $p<0,0001$) et /o/ chez les *bassu* et *secunda*. Le *terza* montre moins d'augmentation du SPR de la parole au chant, surtout pour /e, ɛ, a/. Les spectres moyennés à long terme (LTAS) de l'ensemble du chant, confirment une nette augmentation de l'énergie comprise entre 2000Hz et 6000Hz pour les *bassu* et *secunda* uniquement, toujours de la parole au chant, mais sans pic formantique visible.

3.3 Corrélation avec la fréquence fondamentale

	Bassu	Secunda	Terza		Bassu	Secunda	Terza
Co F0/F1	**0.6**	**0.5**	**0.6**	Co F0/F4	**0.4**	0.03	**0.2**
Co F0/F2	-0.1	0.1	**-0.2**	Co F0/SPR	**0.4**	**0.3**	-0.1
Co F0/F3	**0.6**	0.02	0.1	Nombre	267	282	216

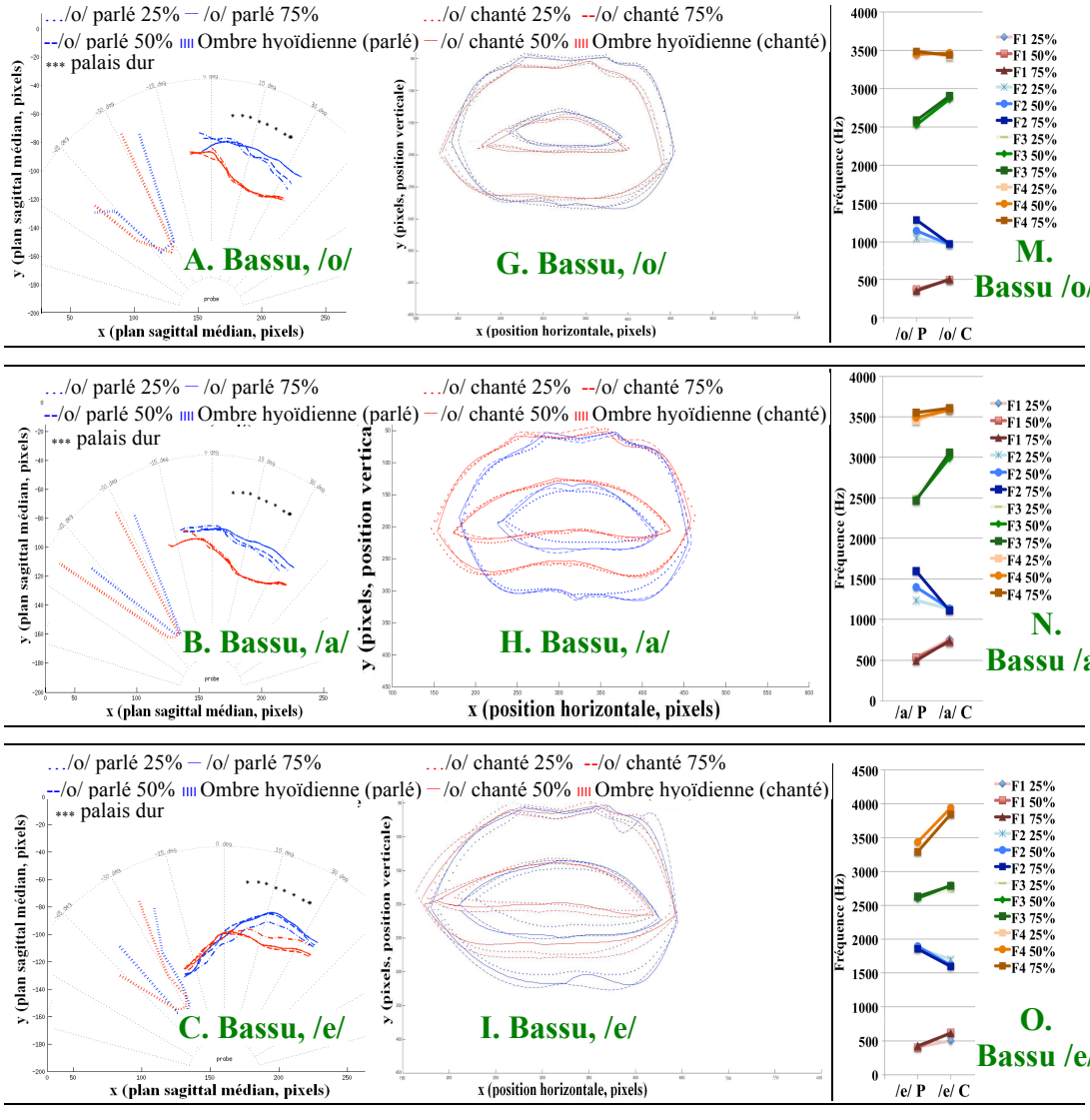
TABLE 2 : Corrélation de Pearson (Co) entre la fréquence fondamentale (F0) des productions parlées et chantées, et les valeurs mesurées (en gras : $p<0,001$)

Il existe une corrélation positive significative entre F0 et F1 pour les données parlées et chantées associées (tableau 2): F1 augmente avec F0, sans autre tendance pour les autres formants et le SPR, sauf une corrélation positive significative pour le *bassu* entre F0 et F3, F4 et le SPR, et entre F0 et

SPR pour le *secunda*. L'analyse des interactions entre formants et harmoniques du spectre de source pourra, dans une recherche ultérieure, compléter l'étude de la relation entre F0 et les formants vocaliques, afin de voir l'existence d'éventuels phénomènes de *formant tuning* (Henrich *et al.* 2011).

3.4 Lien avec les données articulatoires

La figure 4 montre l'ensemble des analyses articulatoires obtenus chez le *bassu* et le *secunda* : visualisation de l'ombre hyoïdienne et du contour lingual à gauche (figure 4, A à F), des lèvres au milieu (figure 4, G à L : légende commune avec les figures A à F : bleu pour les voyelles parlées, rouge pour les voyelles chantées) et fréquences des quatre premiers formants à droite (figure 4, M à R). Les occurrences à 25, 50 et 75% de chaque voyelle sont montrées.



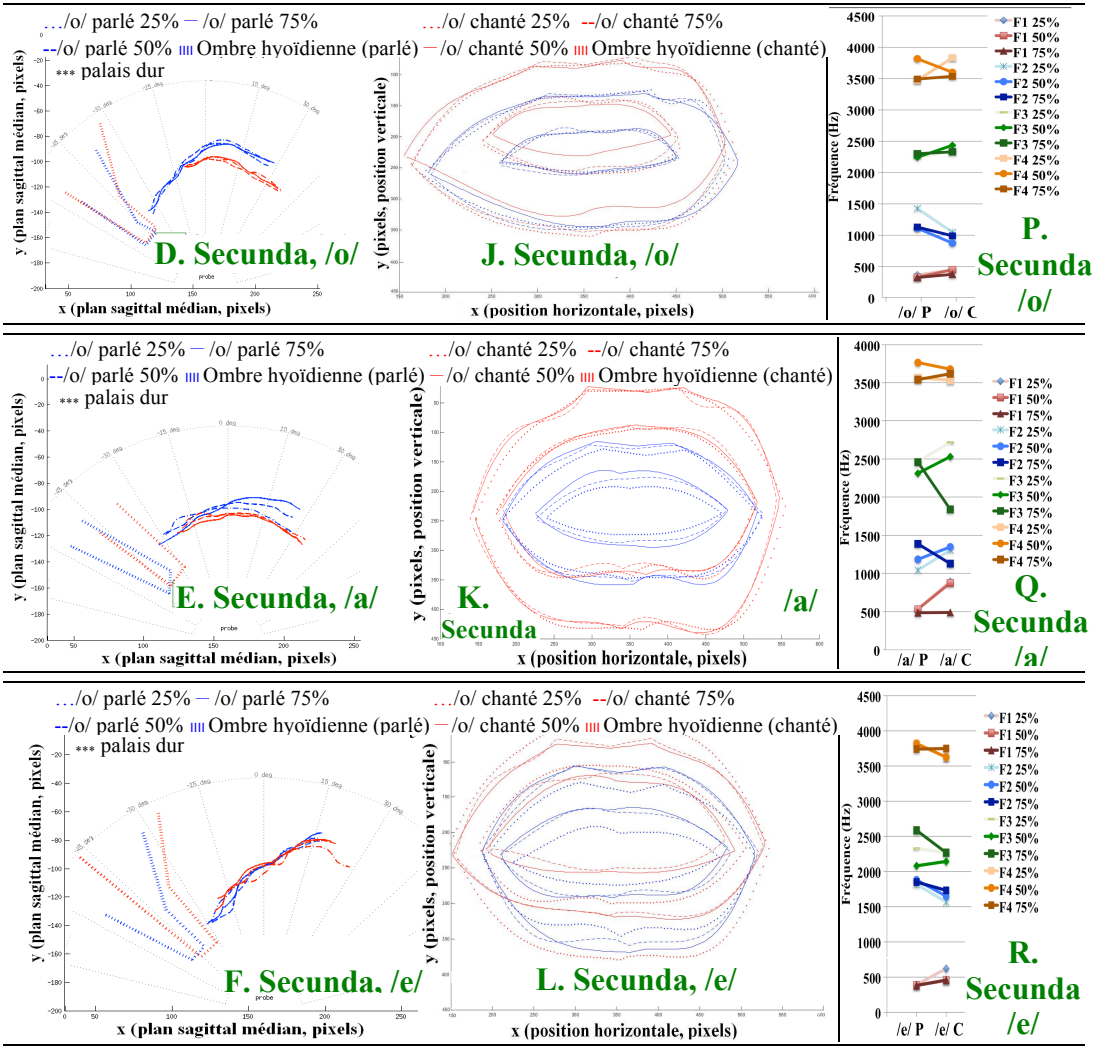


FIGURE 4: position de l'ombre hyoïdienne, de la langue (gauche, A à F) et des lèvres (milieu : G à L) en voix parlée (bleu) et chantée (rouge) pour /o/ de « coghju », le premier /e/ de « pelle », et /a/ de « paghju » par les *bassu* et *secunda*. Droite : formants correspondants (M à R).

De la parole au chant, la forme linguale (figure 4, A à F) s'abaisse pour toutes les voyelles chez les deux sujets, sauf le /e/ du *secunda*. F1 augmente pour ces voyelles (figure 4, M à R). La langue se postériorise pour celles du *bassu*, en même temps qu'on y observe une postériorisation de l'ombre hyoïdienne et un abaissement de F2 (figure 4, M à R). Le *secunda* ne montre pas de postériorisation linguale et hyoïdienne nette de la parole au chant, malgré un abaissement de F2.

L'analyse de la forme des lèvres (figure 4, G à L) montre que toutes sont plus étirées en chant qu'en parole, surtout concernant l'étirement entre l'intérieur des commissures labiales, et surtout pour /a/ produit par le *secunda*. En outre, le *bassu* étire plus à l'intérieur qu'à l'extérieur de ses commissures alors que le *secunda* étire autant l'intérieur que l'extérieur de ses lèvres. Enfin, le *bassu* étire globalement moins ses lèvres que le *secunda* en chant, quelle que soit la voyelle. F3 (figure 4, M à

R) augmente de la parole au chant, sauf à 75% du /a/ produit par le *secunda*. Il existe une légère fermeture des lèvres en chant par rapport à la parole pour le *bassu*. Le *bassu* ferme autant l'intérieur que l'extérieur de ses lèvres alors que le *secunda* ferme légèrement les lèvres à l'intérieur (sauf pour /a/), et les ouvre légèrement à l'extérieur pour /e/ et /o/. F1 augmente de la parole au chant pour l'ensemble de ces productions. La stabilité temporelle de ces paramètres articulatoires et acoustiques (différence entre 25, 50 et 75%) est plus importante en chant qu'en parole, en particulier pour la langue et F2, sauf pour les voyelles et surtout le /a/ du *secunda*.

4 Discussion et conclusion

Nous retrouvons certaines modifications formantiques décrites dans la littérature pour les voyelles chantées par rapport à la parole : F1 plus élevé en corrélation avec l'élévation de F0, abaissement lingual et de F2 (voyelles antérieures), légère élévation de F2 (voyelles postérieures). Cependant, l'augmentation d'aperture supposée se manifester par la nette augmentation de F1 n'a pas pu être démontrée à l'aide de l'ouverture labiale chez le *bassu*, et l'abaissement de F3 en lien avec un arrondissement des voyelles chantées en chant lyrique ne se retrouve pas dans nos données.

Ces modifications formantiques montrent une centralisation de l'espace vocalique, comme l'évoque Lortat-Jacob (1998) pour les chants sardes : « /i/ devient /e/ (…) et /u/ devient /o/ » (p. 129). Cette centralisation se produit plus au niveau postérieur pour le *bassu* que le *secunda* : « ceux qui chantent dans l'aigu poussent [les voyelles] vers l'avant pour les colorer d'aigu », les basses « les tirent vers l'arrière pour les colorer de grave » (Lortat-Jacob, 1998). Pour tous nos chanteurs, la bonne prononciation des voyelles est plus importante que l'adaptation du texte au chant.

Omori *et al.* (1996) obtinrent des valeurs de SPR égales en parole et chant chez des non chanteurs (-22,5dB en moyenne), alors que leurs valeurs passent de -20dB (parlé) à -11,5dB (chanté) chez des amateurs, et de -18dB à -11dB chez des chanteurs lyriques professionnels. Nos moyennes de SPR (-25dB : *bassu*, -35,2dB : *secunda* ; -30dB : *terza* en parole ; -19dB : *bassu*, -24dB : *secunda* ; -26dB : *terza* en chant) sont donc en deçà de valeurs de chanteurs amateurs classiques, et varient d'un chanteur à l'autre. Aucun rapprochement de F3 et F4, aboutissant, d'après la littérature, à la présence du formant du chanteur pour le chant lyrique, n'a été constaté pour nos sujets, peut-être en raison de l'absence d'accompagnement ou des conditions d'enregistrement (chanteurs séparés).

Enfin, il existe une variabilité inter-sujets, probablement voulue par ces chanteurs : chacun possède sa propre façon d'interpréter la *Paghjella* pour en reconnaître l'interprète. Les ornementations de la *terza* rendent aussi ses productions plus variables. Le fait que le *secunda* et la *terza* soient bilingues contrairement au *bassu*, et leurs différents modes d'apprentissage de cette technique vocale, explique aussi la différence de comportement articulatoire entre le *bassu* et le *secunda*. Ces données doivent être obtenues chez un plus grand nombre de chanteurs : les variations observées pourraient provenir de différences anatomiques ou de stratégies de production individuelles différentes. Cette variabilité individuelle est aussi due à une analyse vocalique tous contextes phonétiques confondus. Une étude plus approfondie de ces voyelles considèrera la nature et l'influence de ces contextes, et aussi celle de l'accentuation. Enfin, des tests perceptifs complèteront ces données de production pour mieux comprendre la gestion vocalique dans ces chants.

Remerciements

Ce travail a bénéficié d'une aide de la Commission Européenne au titre du projet *i-Treasures* portant la référence FP7-ICT-2011-9-600676-i-Treasures, ainsi que du LabEx *EFL* (ANR-10-LABX-0083). Nous remercions également les chanteurs pour leur participation à cette étude.

Références

BITHELL C. (2007). *Transported by song : Corsican voices from oral tradition to world stage* (Vol. 5). Lanham, Maryland : Scarecrow Press.

BLOOTHOOFT G., PLOMP R. (1984). Spectral analysis of sung vowels. I. Variation due to differences between vowels, singers, and modes of singing. *J. Acoust. Soc. Am*, 75(4) : 1259-1264.

BOERSMA P. & WEENINK D. (2016). Praat: doing phonetics by computer [Computer program]. Version 6.0.13, retrieved 31 January 2016 from http://www.praat.org/

BURNS P. (1986). Acoustical analysis of the underlying voice differences between two groups of professional singers: opera and country and western. *Laryngoscope*, 96(5): 549-554.

CATINCHI P. J. (1999). *Polyphonies corses*. Cité de la musique. Paris : Actes sud, 150.

CHAWAH P., AL KORK S. K., FUX T., ADDA-DECKER M., AMELOT A., AUDIBERT N., DENBY B., DREYFUS G., JAUMARD-HAKOUN A., PILLOT-LOISEAU C., ROUSSEL P., STONE M., XU K., and CREVIER-BUCHMAN L., An educational platform to capture, visualize and analyze rare singing. Actes de *Interspeech 2014*, Singapour.

DALBERA-STEFANAGGI M.J. (2002). *La langue corse.* Paris : Presses Universitaires de France, Collection « Que sais-je ».

FUX T., AMELOT A., CREVIER-BUCHMAN L., PILLOT-LOISEAU C., ADDA-DECKER M. (2014). MATTONG: une interface graphique sous MatLab pour le suivi du contour de la langue à partir d'images ultrasons. Actes des *Journées d'Etude sur la Parole*, Le Mans, 2014.

HENRICH, N., SMITH, J., WOLFE, J. (2011). Vocal tract resonances in singing: Strategies used by sopranos, altos, tenors, and baritones. *J. Acoust. Soc. Am*, 129(2) : 1024-1035.

GIBSON A. (2010). *Production and perception of vowels in New Zealand Popular Music*. Mémoire de Master, Aukland University of Technology.

LORTAT-JABOB B. (1998). *Chants de Passion, au coeur d'une confrérie de Sardaigne*. Paris: CERF.

NAIR A., NAIR G., REISHOFER G. (2016). The Low Mandible Maneuver and Its Resonential Implications for Elite Singers. *Journal of Voice*, 30(1), 13-32.

OMORI K., KACKER A., CARROLL L.M., RILEY W.D., BLAUGRUND S.M. (1996). Singing Power Ratio: quantitative evaluation of singing voice quality. *Journal of Voice*, 10, 3, 226-235.

PILLOT C., Vaissière J. (2007). Spectral correlates of carrying power in speech and western lyrical singing according to acoustic and phonetic factors. Actes du *15th International Congress of Acoustics*. Madrid, 1-7 septembre 2007, 1-6.

SUNDBERG J. (1987). *The Science of the singing voice*. Dekalb, Illinois : Northern Illinois University Press.

SUNDBERG J., SKOOG J. (1997). Dependance of jaw opening on pitch and vowel in singers. *Journal of Voice*, 11(3), 301-306.

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

TITZE I.R., MAPES S., STORY B. (1994). Acoustics of the tenor high voice. *J. Acoust. Soc. Am*, 95(2) : 1133-1142.

La prosodie du focus dans les parlers algérois et oranais

Ismaël Benali
CLILLAC-ARP, Université Paris Diderot, 8 place Paul Ricoeur75013, Paris, France
ismail.benali@linguist.univ-paris-diderot.fr

RESUME

Le but de cette étude est d'étudier les caractéristiques prosodiques de différents types de focus dans les parlers algérois et oranais.
Il ressort de l'analyse acoustique des productions des locuteurs que les récurrences des schèmes prosodiques qui distinguent les deux parlers sont observées dans deux types de focus : le focus étroit d'insistance quand il est placé à la frontière d'un groupe intonatif et le focus interrogatif. Le premier est réalisé dans le parler algérois par un contour montant descendant. Dans le parler oranais, il est produit par un contour plat ou légèrement montant ou descendant. On retrouve, dans le focus interrogatif, le mêmes contour intonatif plus amplifié du focus d'insistance chez les Algérois alors que chez les Oranais la dernière syllabe est toujours montante précédée d'une descente. Le focus de contraste est produit différemment dans le même dialecte avec plus d'allongement en oranais. La réalisation du focus large n'est pas distinctive.

ABSTRACT

The prosody of focus in Algiers and Oran dialects
The aim of this study is to compare the prosody of focus in Algiers and Oran varieties. Prosodic features associated with different types of focus were examined.
It appears from the acoustical analysis of the speakers' productions that recurrences of prosodic patterns that differentiate the two dialects were observed in only two types of focus: emphatic focus and interrogative focus placed at the final edge of an intonation phrase. The former is produced in the Algiers dialect by a rise-fall contour. In the Oran dialect, this focus is realized with a flat or slightly rising or falling contour.
In the interrogative focus the intonation contours are the same, but more amplified for Algiers speakers; for Oran speakers, the last syllable is always rising and preceded by a falling contour.
The realization of contrastive focus varied across speakers of the same dialect with a lengthening for Oran speakers. Broad focus was realized with similar intonation patterns for both dialects.

MOTS-CLES : variétés dialectales – arabe algérien – focalisation – prosodie
KEYWORDS: dialectal variations – Algerian Arabic – focus - prosody

1 Introduction

Les parlers algérois et oranais sont caractérisés par des accents régionaux qui sont plus aisément reconnaissables au niveau segmental et lexical que suprasegmental pour des auditeurs algériens.
Peu d'études ont été menées sur la prosodie de l'arabe algérien : R. Ait Oumeziane (Aït Oumeziane, 1981) dans sa thèse sur le parler arabe de Constantine en a fourni une description accentuelle et intonative. P. Georgin (Georgin, 1980) a montré certains aspects de l'intonation du parler algérois dans différentes modalités. Il a relevé que l'intonation de ce qu'il appelle « énonciative avec variante » pour la déclarative, est caractérisée par une courbe recto tonale avec une descente sur la dernière syllabe pour la finalité et une montée sur la dernière syllabe pour la continuation majeure.
L'interrogative réside dans le fait qu'une courbe est ascendante sur l'avant dernière syllabe et descendante sur la dernière syllabe. N. Guella (Guella, 1984), dans sa description de l'intonation du parler de Nedroma (près d'Oran), a mis en relief l'intonation de ce parler du point de vue pragmatique en analysant la place de l'accent nucléaire dans l'organisation thème/rhème et dans le focus de contraste. La déclarative dans ce parler selon Guella, se réalise avec un contour descendant sur la fin de la dernière syllabe. La question totale est produite avec une intonation montante précédée dans certains cas par une descente et la question partielle est caractérisée par une intonation descendante.
D'autres recherches ont traité de la position de l'accent : F. Bouhadiba pour le parler d'Oran (Bouhadiba, 1988) et A. Boucherit (Boucherit, 2006) pour le parler d'Alger. Le parler algérien est plutôt considéré comme la plupart des parlers maghrébins comme langue accentuelle. L'accent en arabe algérien n'est pas distinctif. Il est prédictible en fonction du poids et de la position de la syllabe: il porte sur la finale si elle est surlourde ou si elle est la seule lourde dans le mot. Dans les autres cas l'accent porte sur la pénultième.

Dans une étude antérieure (Benali, 2004) rendant compte des spécificités prosodiques des parlers algérois et oranais dans leur identification, il ressortait que les Algérois produisaient plus de variations mélodiques que les Oranais qui eux avaient tendance à produire plus d'allongements syllabiques. Nous avions observé au cours de cette étude que ces caractéristiques prosodiques étaient plus saillantes dans des contextes particuliers d'énonciation : elles se manifestaient plus clairement lorsque le locuteur parlait avec emphase et implication.

Il nous est paru donc nécessaire d'effectuer une description prosodique des deux parlers dans laquelle le point de vue énonciatif est pris en compte.

Les structures prosodiques de ces variétés dialectales ont été étudiées dans le cadre de la problématique posée par la structure informationnelle représentée par la mise en relief par différents procédés de focalisation : le focus large : focalisation sur l'ensemble de l'énoncé, le focus étroit d'insistance : emphase avec insistance sur un élément d'un énoncé, le focus étroit de contraste : emphase sur un élément contrasté d'un énoncé et enfin, le focus interrogatif : emphase d'un élément de l'énoncé sur lequel porte la question. Ex : Jean est **parti** ? (Jean est vraiment parti ?). Le focus contrastif a une fonction corrective Ex : **Jean** est parti (et non Pierre). Le focus d'insistance a plusieurs fonctions: informative, identificatrice, appellative… Ex : Qui est parti ? **Jean** est parti. Il y a insistance dans la mesure où Jean a été évoqué auparavant.

Le focus étroit est marqué dans la plupart des langues par une augmentation des trois paramètres acoustiques : la fréquence fondamentale F0, la durée et l'intensité (Hirst, Di Cristo, 1998).
Le focus de contraste n'est pas toujours marqué prosodiquement comme l'a remarqué 't Hart en néerlandais ('t Hart, 1998). Dans les dialectes du sud de la Suède, le focus étroit ne se distingue pas prosodiquement des énoncés sans focus (Gårding, 1998).

Dans certaines langues, la focalisation est souvent accompagnée par une désaccentuation avant et après le mot focalisé. En danois cet abaissement des syllabes autour de celle qui est focalisée est très important (GrOnnum, 1998). En arabe marocain, seul le mot focalisé est accentué (Benkirane 1998). Dans leur comparaison de trois dialectes arabes (marocain, koweitien et yéménite) Yeou et al (Yeou et al., 2007) ont démontré que ces dialectes partageaient la même stratégie dans la réalisation du focus de contraste qui consiste en un mouvement montant descendant. Les locuteurs marocains se distinguent par une désaccentuation des syllabes qui précédaient le mot focalisé, ce qui n'était pas le cas pour les locuteurs yéménites et koweitiens. Le contour mélodique du focus contrastif était plus localement défini chez les locuteurs yéménites et koweitiens alors qu'il englobait tout le mot focalisé chez les locuteurs marocains. Les locuteurs koweitiens réalisaient en plus du contour montant descendant, un contour montant très élevé à la fin du mot focalisé.
Dans cette étude, les locuteurs marocains se distinguent par un effet significatif de la réalisation de l'alignement selon la structure syllabique : Le pic de F0 se produit à l'intérieur de la syllabe accentuée quand elle est fermée et à l'extérieur quand elle est ouverte. Les locuteurs koweitiens et yéménites produisent ce pic à la fin de la voyelle accentuée que ce soit en syllabe ouverte ou fermée.
En arabe tunisien, le focus affecte positivement aussi bien la durée de la syllabe accentuée que celle de la syllabe non accentuée (Bouchhioua, 2009). Les finales accentuées sont plus allongées et la F0 et l'intensité de la syllabe accentuée augmentent sous l'effet du focus.
Les résultats d'une étude sur l'arabe égyptien de S. Hellmuth (Hellmuth, 2011) ont montré une augmentation de la F0 dans les mots focalisés et une compression de celle-ci dans les mots qui suivent. Cependant, le statut d'information donnée ou nouvelle (givenness) du mot post-focal n'a pas d'effet sur la F0. La durée et l'intensité n'ont subi de variation ni sous l'effet du focus ni sous l'effet du *givenness*.

Les objectifs de notre étude sont donc de comparer le parler algérois et le parler oranais du point de vue prosodique dans la parole lue et spontanée en prenant en compte la structure informationnelle et d'identifier les caractéristiques prosodiques des deux parlers dans les différents types de focus.

2 Méthodologie

2.1 Locuteurs

Les enregistrements ont été réalisés dans les villes d'Oran et d'Alger dans le milieu universitaire. 20 locuteurs algérois (15 hommes et 5 femmes) et 20 locuteurs oranais (10 hommes et 10 femmes) âgés de 22 à 30 ans ayant vécu toute leur enfance et leur adolescence en Algérie ; ont été enregistrés dans des salles au calme à l'aide d'un minidisque et d'un micro-cravate. Ils sont pour la majorité des étudiants de filières différentes.

2.2 Corpus

Le corpus se compose de parole spontanée et de parole lue. Les énoncés de la parole « spontanée » ont été extraits de conversations entre les locuteurs et l'interviewer. Elles étaient imprégnées de sujets établis ou spontanés relatifs à la vie quotidienne et à leurs préoccupations. Une centaine d'énoncés ont été extraits pour chacun des locuteurs.
Les emprunts et l'alternance codique présents très usités dans les parlers urbains ont été conservés pour l'analyse.
Exemple d'une production d'un locuteur oranais : [kæ:jnø demomymɑ̃ tæʕ sbaɲo:l kimæ sɑ̃takru:z]
"Il y a des monuments qui appartiennent aux Espagnols comme Santa-Cruz"

Pour ce qui est de la parole lue, les locuteurs devaient lire 157 énoncés dans le dialecte de l'arabe algérien a priori « standard ». Les différences segmentales et lexicales des parlers algérois et oranais ont été évitées. Les énoncés ont été numérisés avec le programme Sound Forge avec une fréquence d'échantillonnage de 22050 Hz, 16 bits mono. La durée des énoncés varie entre 0,8 et 8 secondes.

2.3 Types de focus

Le but de cette étude étant de décrire la réalisation des différents focus en algérois et oranais, nous avons pour ce faire, soumis aux locuteurs une série de questions à partir d'une phrase de départ qui leur a été présentée graphiquement pour provoquer ces focus.

Exemple:
Phrase de départ "Ali (il) est malade" [ʕali rah mriːdˁ]

a) focus large : Q: "Qu'est-ce qu'il y a?" → R: « **Ali est malade** ». [**ʕali rah mriːdˁ**].

b) focus étroit d'insistance: Q: "Qu'est-ce qu'il a Ali?" → R: "Il est **malade**." [rah **mriːdˁ**].
Q: "Qui est malade?" → R: "**Ali** est malade." [**ʕali** rah mriːdˁ].

c) focus étroit de contraste: Q: "Ali va bien?" → R: "Non, il est **malade**." [(la) rah **mriːdˁ**]
Q: "Mohamed est malade?" → R: "Non, **Ali** est malade." [(la) **ʕali** rah mriːdˁ]
Les réponses des locuteurs pour obtenir les focus d'insistance et de contraste ne consistaient pas à la répétition de la phrase de départ mais à des énoncés plus courts et à l'emploi dans le focus de contraste de l'adverbe de négation [la] chez les oranais et [laela] chez les algérois. Nous avons décidé de garder ces productions parce qu'elles correspondaient plus à la réalité.

d) focus interrogatif: Le focus interrogatif est obtenu par la production de questions totales sans mot interrogatif :
"Ali est **malade** ?" dans le sens de "Ali est vraiment malade ?" [ʕali rah **mriːdˁ** ?]
"**Ali** est malade ?" dans le sens "C'est Ali qui est malade ?" [**ʕali** rah mriːdˁ ?]

2.4 Paramètres prosodiques

Les paramètres prosodiques analysés sont :
a) **la configuration mélodique** : forme du contour mélodique.
b) **le registre tonal** : hauteur moyenne de la F0.
c) **l'étendue tonale** : écart entre les fréquences minimales et maximales.
d) **l'alignement** : association du pic de F0 sur la chaîne segmentale.
e) **la désaccentuation** : phénomène d'abaissement de la courbe mélodique pré/post-focus.
f) **la configuration temporelle** : débit, allongement syllabique.

Les analyses ont été effectués sur le logiciel WinPitch d'analyse de la parole conçu par Philippe Martin (Martin, 2000).

3 Analyse de la parole spontanée

Cette analyse consistait à comparer deux productions pour chaque locuteur algérois et oranais : une avec focus étroit et une autre sans focus étroit.

3.1 Synthèse

De ces analyses, ressortent les observations suivantes : le focus étroit est marqué prosodiquement dans les deux parlers.
Chez les Algérois, il est caractérisé par un pic mélodique sur la pénultième accompagné d'une chute sur la finale et par un contour montant descendant sur les mots monosyllabiques. La chute mélodique est plus importante quand elle est située à la frontière d'un groupe intonatif et elle est marquée par un écart tonal large. L'alignement du pic de F0 se situe généralement avant le noyau focal.
Le phénomène de désaccentuation se produit en dehors du focus dans la plupart des cas.
Le focus large est caractérisé par un registre haut et une baisse de l'intensité.
Chez les Oranais, le focus étroit se réalise par un allongement des syllabes portant l'accent primaire et secondaire. Le contour mélodique est soit montant sur la pénultième soit plat. La finale reçoit un contour bas descendant ou plat. Le pic de F0 se réalise généralement sur le noyau focal ou après celui-ci. La désaccentuation préfocale a été aussi observée.
Le focus large est généralement produit avec un registre plus haut et ne manifeste pas de variations mélodiques.

3.2 Illustrations

Les exemples suivants montrent des productions de focus étroits d'un locuteur algérois (figure 1) et d'un locuteur oranais (figure 2).

La figure 1 représente la production d'un focus étroit par un locuteur algérois
[luka:n tsema insa:n ʕandu **imkaniyæ:t**] « S'ils avaient les **moyens »**
Le locuteur se plaint du manque de moyens des comédiens.
Focus étroit d'insistance sur le mot [**imkaniyæ:t**], contour montant accompagné d'une chute mélodique sur la dernière syllabe. Le pic de F0 (entouré) est aligné avant le noyau focal. Etendue tonale de 8Dt.
Les constituants préfocaux de l'énoncé ne sont pas accentués.

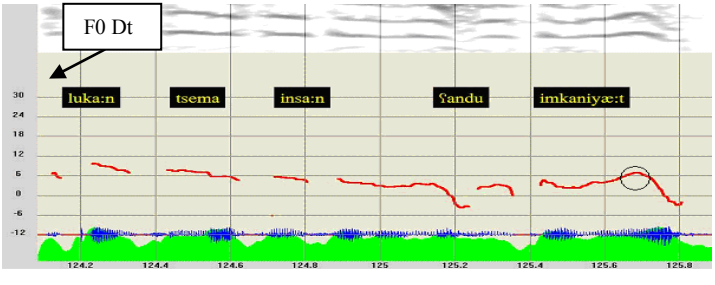


FIGURE 1: Focus étroit dans [imkaniyæ:t] produit par un locuteur algérois

La figure 2 montre une production d'un locuteur ornais :

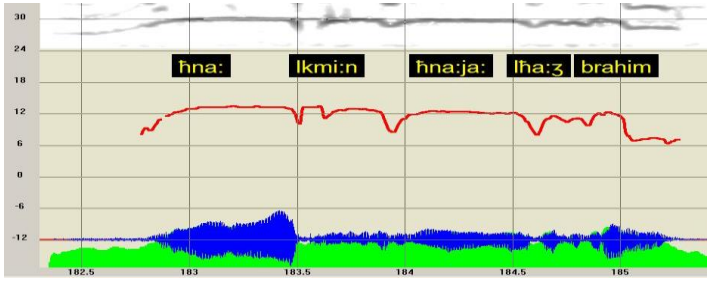


FIGURE 2*:* Focus étroits sur [ħna:], [ħna:ja:] et [brahim] produits par un locuteur oranais

[**ħna**: lkmi:n **ħna:ja:** lħa:ʒ **brahim**] « **Nous** à El Kmin **nous** c'est Haj **Brahim.** »

Le locuteur oranais parle de son quartier El Kmin qui est connu pour avoir enfanté le chanteur Khaled (Hadj Brahim).

Emphase sur le « nous » avec répétition pour annoncer la personnalité de son quartier. Il traduit l'implication du locuteur.

Le deuxième « nous » *ħna:ja:* englobe les habitants de El Kmin qui s'identifient à Hadj Brahim selon le locuteur.

Cette emphase se réalise par l'allongement de toutes les syllabes des deux « nous » *[ħna:]* *[ħna:ja:]* ainsi que de la dernière syllabe du troisième focus **[***brahim***]** avec un abaissement de hauteur. Ils sont tous produits avec un contour plat.

3.3 Conclusion

Ce qui caractérise le parler algérois est le contour montant descendant et surtout la chute mélodique en fin de groupe intonatif dans le focus étroit.

Le parler oranais quant à lui se manifeste par un allongement des syllabes accentuées avec des abaissements de contours qui sont généralement plats.

Quand ces configurations prosodiques sont redondantes dans un même énoncé, elles donnent une impression de rythme caractéristique du parler en question.

4 Analyse de la parole lue

Les résultats suivants sont issus de l'analyse de l'énoncé "Ali (il) est malade" [ʕali rah mriːdˤ]. Les analyses acoustiques sont tirées de la production de 12 locuteurs de chaque parler pour les focus d'insistance et de contraste. 10 locuteurs de chaque parler pour le focus large et 15 locuteurs de chaque parler pour le focus interrogatif.

4.1 Résultats des analyses acoustiques

Ce qui distingue les Algérois des Oranais est la production d'une étendue tonale significativement plus large chez les premiers dans le focus d'insistance (p= 0.010) et interrogatif (p= 0.0009). Les Oranais quant à eux, se distinguent par un débit significativement plus lent dans la production du focus de contraste (p= 0.03). Le focus interrogatif est produit avec une étendue plus large par rapport aux autres focus dans les deux parlers. Les Oranais produisent le focus d'insistance avec une étendue plus large par rapport au focus large. Dans les deux parlers, il est produit avec un débit plus lent et le focus de contraste l'est aussi seulement chez les Oranais. Chez ces derniers, le focus interrogatif est réalisé avec un débit rapide par rapport aux autres focus. Toutes ces données sont significatives.

4.2 Modélisation de la courbe mélodique

Cette modélisation de la courbe mélodique des locuteurs algérois et oranais dans différents types de focus représentée dans la figure 3, a été faite à partir de productions de 15 locuteurs de chaque parler pour l'énoncé :

"Ali est malade." [ʕali rah mri:dˁ.]

Le dernier mot est monosyllabique de type 'CCV:C.

Trois mesures par syllabes ont été prises pour illustrer les variations des contours intonatifs.

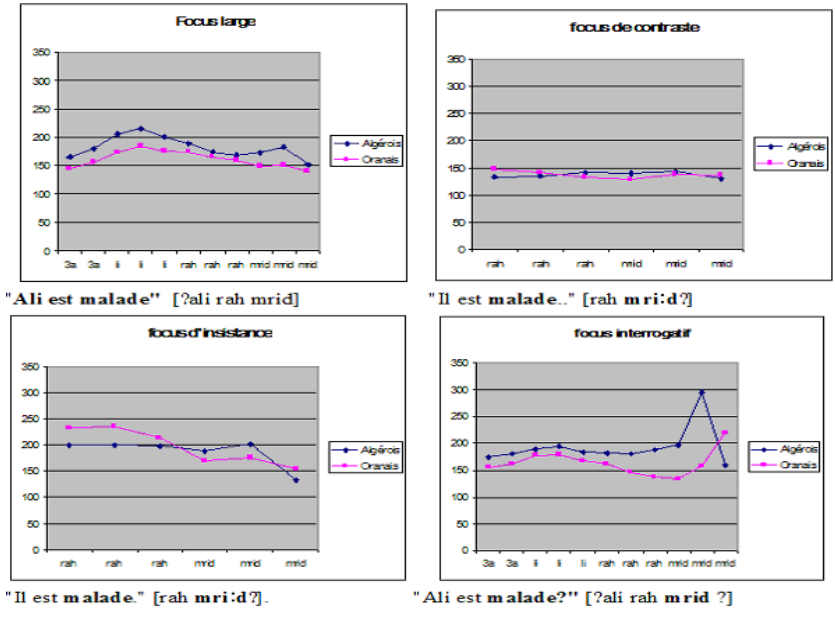


FIGURE 3 : Modélisation de la courbe mélodique des locuteurs algérois et oranais dans différents types de focus

Le focus d'instance est réalisé dans le parler algérois par un contour montant descendant sur la syllabe finale accentuée. Dans le parler oranais, le focus d'insistance est réalisé par un contour plat ou légèrement montant sur la syllabe finale accentuée et montant sur la pénultième accentuée et légèrement descendant sur la finale. Dans le focus interrogatif on retrouve le même contour intonatif du focus d'insistance mais plus amplifiés chez les Algérois alors que chez les Oranais la dernière syllabe est toujours montante précédée d'une descente. On observe qu'il n'y a pas de variations significatives entre l'algérois et l'oranais dans les focus larges et de contraste.

5 Conclusion

Le focus interrogatif et le focus d'insistance ont une prosodie qui distingue mieux les deux parlers. Le focus large et le focus de contraste se réalisent différemment même entre les locuteurs du même parler et leurs prosodies ne permettent pas de les différencier dans la parole lue. Cependant, les Oranais ont tendance à désaccentuer les deux focus et à ralentir leur débit dans le focus de contraste.
Le focus d'insistance se caractérise dans le parler algérois par un contour montant descendant sur le mot focalisé avec une chute mélodique, un allongement de la dernière syllabe et une étendue tonale plus large. Ce contour se réalise sur la dernière ou sur les deux dernières syllabes. Le pic de F0 est aligné avant le noyau focal chez les Algérois. Dans le parler oranais ce focus se réalise généralement par un allongement de la dernière syllabe et par un contour mélodique plat ou légèrement montant ou descendant.
Les portions focalisées en finale deviennent des 'clichés mélodiques' caractéristiques des deux variétés dialectales.

Références

AÏT OUMEZIANE R. (1981). *Le Parler Arabe de Constantine*. Paris, Université Sorbonne Nouvelle Paris 3.

BENALI I. (2004). Le rôle de la prosodie dans l'identification de deux parlers algériens: l'algérois et l'oranais. *Workshop MIDL:* 128-132.

BENKIRANE T. (1998). Intonation in Western Arabic (Morocco). *Intonation systems: a survey of twenty languages*: 345-359.

BOUCHERIT A. (2006). Algiers Arabic. *Encyclopedia of Arabic Language and Linguistics*: 58-66.

BOUCHHIOUA N. (2009). Stress and Accent in Tunisian Arabic. *First International Conference on Intonational Variation in Arabic*.

BOUHADIBA F. A. N. (1988). *Aspects of Algerian Arabic verb phonology and morphology*. University of Reading.

GÅRDING, E (1998). Intonation in Swedish. *Intonation Patterns: A Survey of Twenty Languages*. Cambridge Cambridge UP: 112-130.

GEORGIN P. (1980). *Esquisse phonologique et détermination nominale du parler arabe d'Alger*.

GRONNUM N. (1998). Intonation in Danish. *Intonation Patterns: A Survey of Twenty Languages*. Cambridge Cambridge UP: 131-151.

GUELLA N. (1984). On Syllabication, Stress and Intonation in an Algerian Arabic Dialect. *Etudes et Recherches en Linguistique Oran* (5): 1-19.

HART J. t. (1998). Intonation in Dutch. *Intonation Systems*: 96.

HELLMUTH S. (2011). Acoustic cues to focus and givenness in Egyptian Arabic. *Instrumental Studies in Arabic Phonetics* 319: 301.

HIRST D. and DI CRISTO A. (1998). *Intonation systems: a survey of twenty languages*: Cambridge University Press.

MARTIN P. (2000). WinPitch 2000. a tool for experimental phonology and intonation research. *Proceedings of the Prosody 2000 Workshop*.

YEOU M., EMBARKI M., AL-MAQTARI S. (2007). Contrastive focus and F0 patterns in three Arabic dialects. *Nouveaux cahiers de linguistique française*: 317.

Que disent nos silences ? Apport des données acoustiques, articulatoires et physiologiques pour l'étude des pauses silencieuses

Muriel Lalain[1], Thierry Legou[1], Camille Fauth [2], Fabrice Hirsch[3], Ivana Didirkova[3]

(1) Université d'Aix-Marseille, CNRS, LPL, UMR 7309, Aix-en-Provence, France
(2) Université de Strasbourg, Institut de Phonétique, E.A. 1339 LILPa, Strasbourg, France
(3) Université Paul Valery, Montpellier 3, CNRS, Praxiling, UMR 5267, Montpellier, France
muriel.lalain@lpl-aix.fr

RESUME

Si la rhétorique s'est intéressée très tôt à la pause, il a fallu attendre le XXème siècle pour que d'autres disciplines – la psycholinguistique, le traitement automatique des langues, la phonétique – accordent à ces moments de silence l'intérêt qu'ils méritent. Il a ainsi été montré que ces ruptures dans le signal acoustique, loin de signer une absence d'activité, constituaient en réalité le lieu d'une activité physiologique (la respiration) et/ou cognitive (planification du discours) qui participent tout autant au message que la parole elle-même.

Dans cette étude pilote, nous proposons des observations et des pistes de réflexions à partir de l'analyse des pauses silencieuses dans un corpus de parole lue et semi dirigée. Nous mettons notamment en évidence l'apport de l'analyse conjointe de données acoustiques, articulatoires (EMA) et physiologiques (respiratoires) pour l'identification, parmi les pauses silencieuses, des pauses respiratoires, syntaxiques et d'hésitation.

ABSTRACT

What do our silences say? Contribution of acoustic, articulatory and physiological data to the study on silent pauses.

While rhetoric has been interested in pauses since a long time, researches on those brief moments of silence in other scientific fields, such as psycholinguistics, natural language processing or phonetics were rather rare before the 20th century. Thus, it has been shown that ruptures in acoustic signal are not a sign of a lack of activity; they are in fact the place of a physiological (respiration) and / or cognitive (discourse planning) activity. Both are as important for the message as the speech activity.

In this pilot study, observations and reflections are made, based on analysis of silent pauses in semi-directed and read speech. Our research shows the importance of considering the acoustic, articulatory (EMA) and physiological (Biopac) data together, in order to identify respiratory, syntactic and hesitation pauses within the brief silences.

MOTS-CLES : pauses ; mouvements articulatoires ; EMA ; Biopac ; respiration ; déglutition

KEYWORDS: pauses ; articulatory movements ; EMA ; Biopac ; respiration ; deglutition

1 Introduction

La pause fait partie intégrante des tours de parole. A ce titre, elle a toujours suscité un certain intérêt depuis l'Antiquité. Depuis le milieu du XXème siècle, linguistes et psycholinguistes ont également mené une réflexion sur ces événements présents dans la parole. Différents travaux ont ainsi rendu compte de la quantité des pauses dans le discours, leur durée, leur nature et leurs fonctions. Sur le plan phonétique, deux types de pauses peuvent être relevés : les pauses pleines et les pauses vides. Les premières correspondent à des allongements de sons, à la réalisation d'un *schwa* (Maclay et Osgood, 1959) ou à l'utilisation d'un autre type de filler (« hum » ou « bin »). Les pauses vides, elles, peuvent être définies comme une interruption du flux de parole se répercutant sur le signal acoustique par une amplitude nulle ou non-significative (Duez, 2003).

D'après Goldman-Eisler (1968) les pauses constituent près de 50% du temps de parole lors d'une description d'image. Ce bref moment de silence est ainsi le signe d'une activité cognitive importante puisque la pause permet de reprendre sa respiration (pauses physiologiques), de planifier le contenu de son message et structurer son énoncé (pauses syntaxiques), et de procéder à une recherche et sélection lexicale (pauses d'hésitation). Grosjean & Deschamps (1972) ont pour leur part montré que les pauses d'hésitation étaient moins fréquentes que les pauses syntaxiques, mais ce résultat est à relativiser en fonction de la situation de parole (lecture *vs* description) et du locuteur lui-même, selon sa plus ou moins grande habileté en lecture par exemple (Lalain et al, 2014). La durée des pauses vides a également suscité un certain intérêt. Grosjean & Deschamps (1975) ont montré qu'elles durent en moyenne 520 ms en parole spontanée et 1320 ms lorsqu'il s'agit de décrire une image. Ces résultats, tout comme ceux de Duez (2001) ou de Goldman et al (2010) montrent que la durée de la pause dépend du style de parole (lecture, parole conversationnelle,…). Cette dernière étude a également révélé que l'étendue de la pause était fonction de la présence ou non d'une prise de respiration et/ou d'une déglutition de salive, tandis que l'étude de Hirsch *et al.* (2015) a montré que les caractéristiques temporelles des pauses peuvent également dépendre de la charge émotionnelle et de la thématique abordée. Les durées des pauses peuvent également dépendre de l'activité motrice qui s'y déroule : l'étude articulatoire et perceptive d'Abry *et al.* (1996) a ainsi montré que des gestes anticipatoires d'arrondissement des lèvres ou d'abaissement de la mandibule pouvaient être présents dans les pauses, mais ce phénomène serait partiellement lié au locuteur et à la durée de la pause.

Ainsi, les travaux sur la pause ont permis de mieux connaître leur nature (syntaxique, respiratoire, recherche lexicale, focalisation…), leur distribution ou encore leur durée, et ce, en fonction de différents styles de parole. Il est ainsi admis en règle générale deux grands types de pauses, et ce dans tous les styles de parole : les pauses syntaxiques et les pauses d'hésitation dont on sait qu'elles se distinguent à la fois par leur fréquence d'apparition, leur durée et leur localisation syntaxique. Cependant, les interruptions sonores d'ordre syntaxique recouvrent également les pauses dont la fonction peut être physiologique, comme les pauses respiratoires ou celles dévolues à la déglutition.

Notre objectif, avec cette étude pilote, est de prolonger les recherches menées sur la pause en élaborant un protocole expérimental fiable, permettant d'utiliser des données à la fois acoustiques, articulatoires et physiologiques pour l'identification objective des différents types de pauses suscités. En d'autres termes, ce protocole visera à mettre en évidence d'éventuelles différences acoustiques, articulatoires et / ou physiologiques entre les différentes catégories de pauses décrites *supra*.

2 Méthodologie

2.1 Recueil des données

Pour cette étude pilote, nous avons constitué un corpus de parole auprès d'une locutrice volontaire au Laboratoire Parole et Langage à Aix-en-Provence. Celle-ci ne présentait ni trouble de l'audition ni trouble de la parole ou du langage. Nous avons recueilli des données acoustiques, articulatoires et physiologiques afin de pouvoir utiliser ces différents indices dans la caractérisation des pauses.

Données Audio, Articulatoires (EMA) et Physiologiques (Biopac)

Les enregistrements audio et articulatoires ont été réalisés avec un articulographe AG501 (5D) à une fréquence de 250 Hz et avec une précision meilleure que le mm sur les trois axes (X, Y, Z). Les huit capteurs utilisés ont été plongés dans du latex avant d'être imprégnés de colle et positionnés comme suit : un capteur sur la mastoïde gauche (1), un sur la lèvre supérieure (2), un sur la lèvre inférieure (3), un à l'inter-incisives supérieure (4), un à l'inter-incisive inférieure (5), deux sur la langue (dos 6 et apex 7), un au niveau du larynx (8). Le capteur placé au niveau du larynx a été collé sur la peau après repérage de la position haute de la partie saillante du larynx lors d'une déglutition. Les capteurs 1 et 4 sont des capteurs de référence permettant la correction des mouvements de tête, les autres capteurs fournissent des indications sur la position et les mouvements des articulateurs. Une ceinture thoracique du type SS5LP connectée à la station d'acquisition MP35 de la société Biopac a été utilisée pour la visualisation de l'activité respiratoire par la variation du périmètre de la cage thoracique.

Synchronisation des enregistrements

L'AG501 intègre un module de synchronisation (SyBox) qui assure d'emblée la synchronisation entre l'enregistrement audio du locuteur et celui des signaux des capteurs de positions. Pour synchroniser l'articulographe avec le Biopac, nous avons développé un module à base d'un microcontrôleur (Arduino). Lorsqu'il reçoit un signal de la SyBox, ce module génère un bip audio (500Hz pendant 500ms) et un signal de déclenchement à destination du Biopac. Le Biopac est configuré pour « attendre » le signal du microcontrôleur avant de commencer l'acquisition. Les données articulatoires, audio et respiratoires sont recueillies à différentes fréquences mais sont ainsi synchronisées.

Corpus

Le sujet devait accomplir 7 tâches (T1 à T7). Les consignes étaient présentées sur un écran d'ordinateur (ppt). Les T1 et T2 consistaient à lire à haute voix et à un rythme usuel *La chèvre de Monsieur Seguin* et un texte intitulé *Pascal*. Ces deux tâches de lecture ont été réitérées en fin d'expérimentation (T6 et T7) dans le but de vérifier un éventuel effet d'habituation au dispositif EMA. Deux tâches de description d'images suivies d'une discussion informelle avec l'un des expérimentateurs ont également été proposées. Il s'agissait de décrire La Une de Charlie Hebdo qui a suivi les attentats du 13 novembre (T3) et des photos de Robert Doisneau (T4) représentant des scènes quotidiennes à Paris. La discussion initiée par l'expérimentateur était centrée sur les ambiances terreur/quiétude représentées par ces différents supports. L'objectif des T3 et T4 était d'obtenir une parole semi contrôlée (description) et conversationnelle (discussion) à l'inverse des T1 et T2 où la production de parole est contrôlée.

2.2 Traitement des données

Les données recueillies au cours de cette étude pilote constituent un corpus d'une durée d'environ 20 mn. Les tâches de lecture, description/discussion et déglutition permettent l'observation de divers phénomènes parmi lesquels la fréquence d'occurrence des pauses, le type de pause et leur durée.

Identification des pauses et de leur nature

L'ensemble du corpus a été transcrit orthographiquement (sous Praat), puis segmenté et annoté de manière semi-automatique (Easy Align). Seules les tiers syllabes et ortho ont été conservées du traitement fait à l'aide d'Easy Align. Les pauses ont été annotées dans la tier Syllabes ; une tier scope a été ajoutée pour l'annotation de segments comprenant la pause ainsi que les syllabes pré et post-pausales.

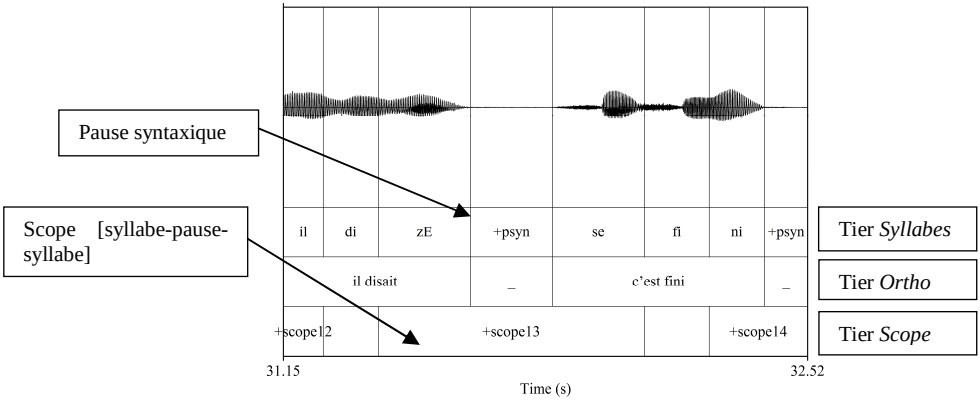


Figure 1 : Etiquetage du signal acoustique avec l'annotation d'une pause syntaxique et d'un scope.

Cette analyse concerne uniquement les pauses vides présentes dans les tâches 1, 2, 3, 4, 6 et 7. Ces pauses ont été identifiées à l'aide d'indices perceptivo-visuels : elles correspondent à un silence perceptible accompagné d'une rupture d'activité acoustique significative visible sur le signal de parole (*cf.* Fig. 1). Aucun seuil de durée n'a été appliqué, contrairement à ce qui est habituellement opéré : l'un des objectifs de cette étude pilote était de vérifier l'apport des données acoustiques articulatoires et physiologiques dans l'identification et la caractérisation des pauses. Seuls ont été éliminés les silences correspondant aux tenues des occlusives sourdes et aux occlusives glottales. La durée de la tenue des occlusives sourdes en position initiale ou post-pausale a été déterminée à partir de la durée moyenne de la tenue de l'ensemble des occlusives sourdes en contexte VCV. Les pauses vides syntaxiques ont été prédites à partir de la fonction POS de Marsatag (pour une description plus complète de cet outil d'enrichissement de données textuelles et de transcriptions de l'oral, voir Rauzy et al., 2014) ainsi que de la ponctuation pour les textes lus. Nous avons ainsi pu identifier, parmi les pauses vides, celles dont la fonction était syntaxique (+psyn) et celles qui relevaient d'autocorrections (+pac) et de cas d'erreurs de décodage (+pdec) en lecture ou d'hésitations avec (+prhes) ou sans (+phes) activité respiratoire en description/discussion. L'examen des signaux de respiration correspondant à chacune des +psyn a permis d'identifier de manière objective les pauses silencieuses respiratoires recodées +prsyn. La figure 2 représente les différents graphes à partir desquels nous avons déterminé la nature des différentes pauses et basé nos observations. Pour

chaque pause, nous avons généré une figure pour l'analyse des mouvements articulatoires selon l'axe Z (élévation-abaissement) et pour l'analyse du signal de respiration. De haut en bas, les graphes 1 à 4 représentent les positions des articulateurs mâchoire/apex/dos /larynx, en Z ; Sur l'axe vertical est représentée l'amplitude du mouvement, le point 0 correspondant à la position de l'articulateur en début de pause. Le graphe 5 rapporte le signal délivré par la ceinture permettant d'observer l'augmentation du périmètre thoracique, le graphe 6 représente le signal audio enregistré. Tous les graphes correspondent à la même durée d'observation, laquelle correspond à la durée des pauses, représentées en secondes sur l'axe commun à tous les graphes au bas de la figure.

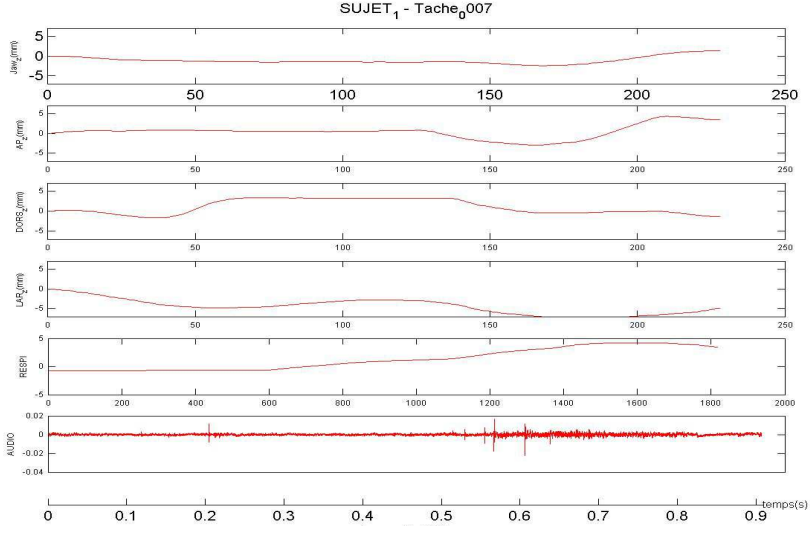


Figure 2 : Graphes utilisés pour la caractérisation des pauses et l'identification de patterns articulatoires et physiologiques ; Ici une pause syntaxique respiratoire (augmentation du périmètre thoracique) précédant le segment [ʒ] (élévation de la mâchoire et de l'apex)

L'utilisation des données acoustiques, syntaxiques et physiologiques nous a ainsi permis d'identifier, parmi les pauses vides initialement relevées, des sous catégories selon les tâches : les pauses vides syntaxiques (+psyn), les pauses vides correspondant à des autocorrections (+pac) ou des erreurs de décodage (+pdec) les pauses vides d'hésitation avec (+prhes) ou sans (+phes) activité respiratoire et les pauses vides syntaxiques respiratoires (+prsyn).

Fréquence d'apparition et durée des pauses selon leur nature et par tâche

Après avoir identifié les différents types de pauses, nous avons examiné leur fréquence d'apparition, à partir de la tier syll qui comprend le codage +p*. Les différentes natures des tâches accomplies par le sujet (lecture vs description/discussion) ont eu pour conséquence la production de parole de durées différentes : les deux tâches de lecture et la tâche de description ne comprennent pas le même nombre de syllabes produites (T1 = 277 syllabes *vs* T2 = 110 syllabes *vs* T3= 240). Afin de pouvoir comparer le nombre de pauses vides entre ces différents exercices de parole, nous avons calculé leur probabilité d'apparition pour chaque tâche : nbpause/(nbsyll+nbpause).

3 Observations préliminaires

3.1 Fréquences d'occurrences des pauses silencieuses

La figure 3 donne les fréquences d'occurrence des différents types de pauses relevés dans chaque tâche. Les pauses silencieuses syntaxiques et les pauses silencieuses syntaxiques respiratoires sont relevées dans les tâches 1 à 7. Les pauses d'autocorrection et de décodage seulement en T6 et T7, et les pauses d'hésitation avec (prhes) et sans (phes) activité respiratoire seulement en T3 et T4.

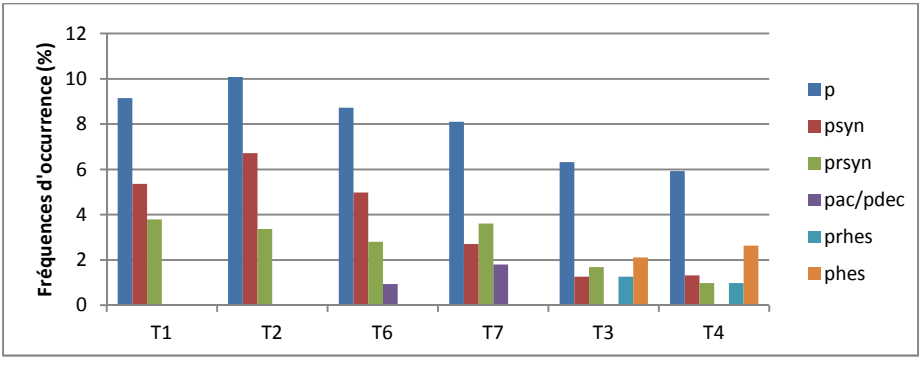


Figure 3 : Fréquences d'apparition (en %) des différents types de pause par tâche

La fréquence d'apparition des pauses syntaxiques est supérieure à celle des pauses respiratoires dans les tâches de lecture alors que cette tendance est inversée en T3et T4. Ceci peut s'expliquer par le fait qu'en lecture, les pauses syntaxiques sont fortement contraintes, notamment en partie par la ponctuation. Concernant les comparaisons T1 *vs* T6 t T2 *vs* T7 rien n'indique une habituation au dispositif EMA, d'autant qu'il n'y a pas de différence en ce qui concerne le débit et la vitesse articulatoire

3.2 Durées des pauses

L'intervalle +ps a été utilisé pour calculer la durée des pauses silencieuses dans chaque tâche. Les durées moyennes des pauses sont données pour chaque type de pause pour les 6 tâches, en secondes.

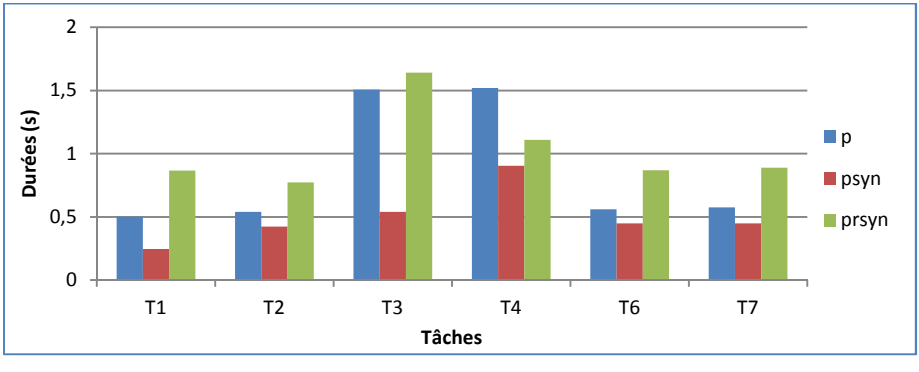


Figure 5 : Durée moyenne (s) des pauses par type de pause et par tâche

On retrouve une observation connue dans la littérature : les pauses sont plus longues en T3 et T4 (tâches semi contrôlées) que dans les autres tâches. Cette tâche de génération de parole implique en effet une charge cognitive plus intense que l'activité de lecture, ce qui se traduit davantage par une durée plus importante qu'une fréquence d'apparition plus élevée. Cet examen de la durée montre également une différence de durée des pauses pour chacune des cinq tâches, entre les pauses syntaxiques et les pauses syntaxiques respiratoires. Ces dernières sont nettement plus longues.

3.3 Patterns physiologiques et articulatoires

A partir des distinctions faites concernant les types de pauses, nous nous sommes intéressés aux articulateurs, en particulier la mandibule, la langue (apex et dos) ainsi que le larynx, sur l'axe Z (plan vertical), et ce afin de repérer des patterns articulatoires récurrents correspondants aux pauses silencieuses respiratoires et aux pauses silencieuses syntaxiques. Cette étape est un préalable à une réflexion concernant le développement d'une méthode d'analyse permettant l'objectivation de ces possibles patterns.

Le capteur 8 positionné sur la peau du cou au niveau du cartilage thyroïde lorsque celui-ci est dans sa position haute en déglutition permet de repérer l'ascension du larynx pendant le deuxième temps de la déglutition. Cette montée du larynx s'accompagne d'une élévation de l'apex et du dos de la langue qui correspondent à l'appui lingual antéro postérieur observable en fin de temps buccal et qui permet la propulsion de la salive vers le pharynx.Ce pattern articulatoire semble récurrent en lecture et en description/narration pour signer les déglutitions physiologiques, lesquelles ont lieu au cours des pauses syntaxiques, en l'absence de respiration (puisque la déglutition s'accompagne d'une inhibition respiratoire pendant les temps oral et pharyngé). Cependant, cette mesure à l'aide du capteur 8 demande à être validée par une étude spécifique visant à préciser la fiabilité de ce capteur et de son positionnement.

A partir des intervalles +scope correspondant à la dernière syllabe précédant la pause et la première suivant la pause (syll +ps syll), nous avons récupéré les mouvements articulatoires pré et post pausaux, afin de pouvoir repérer d'éventuels phénomènes d'anticipation articulatoire pouvant être observés pendant les pauses. Nous avons ainsi observé des patterns articulatoires associés aux pauses silencieuses respiratoires qui semblent refléter les gestes d'anticipation du segment suivant. On peut par exemple relever en fin de pause le mouvement d'élévation de la mandibule et de l'apex que l'on peut attribuer à la préparation articulatoire de la consonne fricative [ʒ] (*cf* Figure 2)

4 Conclusion

L'objectif de cette étude pilote était de montrer la pertinence de l'utilisation conjointe de données acoustiques, articulatoires et physiologiques pour l'identification et la caractérisation des pauses silencieuses dans différents styles de parole. Les observations préliminaires présentées ici ont mis en évidence l'intérêt de l'apport des données physiologiques dans l'identification des pauses silencieuses respiratoires. En effet, on a pu observer que toutes les pauses syntaxiques n'ont pas un rôle physiologique respiratoire et que cette distinction permet de mettre en évidence des caractéristiques temporelles jusqu'ici ignorées : les pauses respiratoires sont en effet plus longues que les pauses syntaxiques. Les données articulatoires semblent pouvoir également contribuer à l'identification d'indices robustes pour caractériser la déglutition (élévation de l'apex, du dos et du larynx). De plus, ces données pourraient contribuer à distinguer les pauses silencieuses syntaxiques et d'hésitation, par des patterns articulatoires différents (anticipation ou recherche articulatoire).

Ces premières observations nous conduisent donc à considérer notre protocole expérimental comme valide, même si quelques améliorations sont à envisager avant le recueil de données auprès d'une plus large population. Pour les tâches de lecture et de discussion, une pré-tâche permettrait d'éviter les hésitations en lecture et les difficultés dans la discussion sur les images. Pour aller plus loin dans l'identification des différents types de pauses (notamment syntaxiques et d'hésitation), une utilisation d'autres fonctions de l'aligneur syntaxique Marsatag pourra s'avérer utile. Le positionnement du capteur 8 au niveau du larynx, devrait faire l'objet d'une étude particulière visant à valider son degré de précision ; cela étant, les observations faites sur ce premier jeu de données sont encourageantes. Enfin, un travail de recherche et développement devrait également être centré sur l'obtention de volumes d'air inspirés et expirés corrélés aux variations de volume de la cage thoracique enregistrés par le Biopac. Ce travail pourrait être conduit avec des mesures associées de Biopac et d'EVA.

Remerciements

Cette recherche a été en partie financée par un IdEx Attractivité de l'Université de Strasbourg « Arythmique » attribué à Camille Fauth, ainsi que par le projet SYNABE (Défi 2016 : Instrumentation aux limites), porté par Fabrice Hirsch.

Références

ABRY C., CATHIARD M.A., EL ABED R., LALLOUACHE M.T., LEROY M.C., PERRIER P, POVEDA F., SAVARIAUX C. (1996). Silent speech production: anticipatory behaviour for 2 out of the 3 main vowel gestures/features while pausing. *First ETRW on speech production modeling*, Autrans, France Mai 1996, p 101-104.

DUEZ D., (2001). Caractéristiques acoustiques et phonétiques des pauses remplies dans la conversation en français. Travaux Interdisciplinaires du Laboratoire Parole et Langage, vol. 20, p 31-48.

DUEZ D. (2003). Le pouvoir du silence et le silence du pouvoir : comment interpréter le discours politique, *MediaMorphoses*, 8, p. 77-82.

GOLDMAN-EISLER F. (1968). *Psycholinguistics. Experiments in spontaneous speech*, New York, Academic Press, 169 p.

GOLDMAN J.-P., FRANÇOIS T., ROEKHAUT S., SIMON A.-C. (2010). Étude statistique de la durée pausale dans différents styles de parole, *Actes des XXVIIIèmes Journées d'Etude sur la Parole, Mons*, 25-28 mai, p.161-164.

GROSJEAN F., DESCHAMPS A., (1972). Analyse des variables temporelles du français spontané, Phonetica, 26, 3, 129-156.

GROSJEAN François & DESCHAMPS Alain, 1975, « Analyse contrastive des variables temporelles de l'anglais et du français », *Phonetica*, 31, 144-184.

HIRSCH F., PEREA F., STEUCKARDT A., VERINE B. (2015). La rédemption est dans la pause. Émotion et prosodie dans l'interview de DSK au vingt heures du 18 septembre 2011, in *Comment les médias parlent des émotions. L'affaire Nafissatou Diallo contre Dominique Strauss-Kahn*, in A. Rabatel, M. Monte et M. Das Graças Soares Rodrigues (Eds), Lambert Lucas, p. 177-194.

LALAIN M., MENDONCA-ALVES L., ESPESSER R., GHIO A., DE LOOZE C., REIS C., (2014). Prosody and reading : temporal and melodic particuliarities in dyslexic child in reading and narration. Rev. Laryngol. Otol. Rhinol, 135, 1, p. 71-82

MACLAY H., OSGOOD C. (1959). Hesitation Phenomena in Spontaneous English Speech, *Word*, 15, p. 19-44.

Que nous apprennent les gros corpus sur l'harmonie vocalique en français ?

Giuseppina TURCO[1], Cécile FOUGERON[1], Nicolas AUDIBERT[1]
(1) Université Sorbonne-Nouvelle Paris 3 - Laboratoire de Phonétique et Phonologie (LPP), UMR 7018, 19 rue des Bernardins, Paris, France
{giuseppina.turco; cecile.fougeron; nicolas.audibert}@univ-paris3.fr

RESUME

Afin de mieux identifier le poids relatif des différents facteurs décrits dans la littérature comme influençant le phénomène d'harmonie vocalique (HV) en français, 33k mots extraits de deux corpus de parole continue et présentant un contexte d'HV possible $V_1C(C)V_2$ ($V_1 \in \{e,\varepsilon,o,ɔ\}$) sont analysés. Le degré d'HV est mesuré en termes d'abaissement du F1 de V_1 induit par la présence d'une V_2 /+haut/ (fermée ou mi-fermée) par rapport à une V_2 /-haut/ (ouverte ou mi-ouverte). Les résultats montrent une HV plus importante pour les voyelles moyennes postérieures que pour les antérieures, et plus faible lorsque l'orthographe favorise une prononciation mi-fermée de V1. Comme attendu, l'HV est plus forte quand V1 est séparé de V2 par une consonne labiale vs. linguale ou par un cluster consonantique sous-jacent vs. un cluster résultant de la chute d'un schwa. En revanche, le style de parole (conversationnelle vs. journalistique) a un effet plus nuancé que celui attendu.

ABSTRACT

What can we learn from big speech corpora about French vowel harmony?

This study sets to determine those highly debated - as yet poorly investigated - factors that affect French vowel harmony (VH). The analyses are based on 33k words extracted from two corpora (journalistic vs. casual speech) with possible $V_1C(C)V_2$ (V_1 {e,ɛ,o,ɔ}) harmonic contexts. VH is measured as the lowering of the first formant of the target V_1 in relation to the V_2 trigger /+high/ (i.e. high and mid-high) vs. /-high/ (i.e. low and mid-low) V2. Results show that VH is stronger for mid-back target vowels than for mid-high vowels and when the realization of the target vowel is not driven by orthographic interference. As expected, VH is stronger with an intervening labial consonant (compared to a lingual consonant) and with consonant clusters without underlying schwa. Speech style is found to have a more subtle effect that the one described in the literature.

MOTS-CLÉS : harmonie vocalique, coarticulation, style de parole, gros corpus
KEYWORDS: vowel harmony, V-to-V coarticulation, speech style, corpus-based study

1 Introduction

Les variations de timbre des voyelles moyennes du français répondent à des influences diverses (variantes régionales ou individuelles, contraintes de distribution en fonction de la structure syllabique, analogie au sein des paradigmes, orthographe, position dans le mot, etc.). Ici, nous nous

intéressons à la modification de timbre que peuvent subir ces voyelles moyennes, lorsqu'elles sont en position non-finale, en fonction de l'aperture de la voyelle finale non-adjacente suivante. Par exemple, dans le paradigme *aime/aimait/aimer*, la première voyelle aura tendance à être réalisée mi-ouverte dans les deux premiers cas ([ɛm], [ɛmɛ]) mais mi-fermée dans le dernier [eme] sous l'influence de la voyelle finale. Décrit en phonologie comme un processus optionnel d'harmonie vocalique (HV) ou de métaphonie, avec transfert à distance de la spécification d'aperture de la source à la cible précédente, cette influence non-locale peut aussi être considérée comme un cas de coarticulation de voyelle à voyelle avec chevauchement des gestes articulatoires des deux voyelles co-produites (Ohala 1994, Fagyal et al., 2003, Nguyen et Fagyal 2008). Si la nature du processus et la question de savoir s'il résulte d'une planification ou d'une articulation anticipée de la voyelle finale est importante (cf. Tilsen 2007), nous n'aborderons cette question qu'indirectement dans cette étude. Notre objectif principal est de mieux comprendre ce phénomène, à la fois bien connu mais relativement mal défini dans la littérature et peu décrit empiriquement, en examinant son occurrence dans de gros corpus de parole continue en français. En d'autres termes, nous ne cherchons pas directement à savoir ***pourquoi*** ce phénomène apparaît mais ***où*** il apparaît (entre quelles voyelles), et puisqu'il est considéré comme optionnel, ***quels facteurs*** influencent son occurrence.

Cette étude s'inscrit donc dans la lignée des études de Fagyal, Nguyen, Boula de Mareüil (2003) et Nguyen & Fagyal (2008) qui, sur un corpus produit en laboratoire, ont étudié ce processus dans les productions de six locuteurs du français standard vs. méridional. Leur étude et leurs résultats montrent à quel point un état des lieux est nécessaire pour comprendre ce phénomène, auquel nous référons dans la suite en termes d'harmonie vocalique, qui est largement admis dans la phonologie du français, mais dont la définition n'est pas totalement claire. En effet, si l'HV est mentionnée dans la plupart des précis de prononciation ou descriptions phonologiques du français (Grammont 1926, Fouché 1959, Dell 1973, Tranel 1989, etc.) comme un processus anticipatoire, optionnel, affectant l'aperture de voyelles moyennes en position non-finale (ou non-accentuée) en fonction de l'aperture de la voyelle finale (accentuée) de la syllabe suivante (i.e. non adjacente), les conditions exactes de son application ne sont pas toujours décrites de la même façon (voir aussi la revue de Fagyal et al., 2003). Pour cela, l'examen de gros corpus de parole continue alignés automatiquement nous semble un paradigme approprié pour examiner plus systématiquement le phénomène dans des productions naturelles, incluant des locuteurs et des mots variés. Les tendances que nous arriverons à mettre à jour nous serviront dans un deuxième temps à construire un corpus plus contrôlé.

La première incertitude dans la définition des conditions d'application de l'HV en français concerne ***la nature de la cible (V_1)***, c'est à dire quelles voyelles moyennes peuvent subir l'harmonie dans un contexte /(#)V_1C(C)V_2#/. Chez Fouché (1959) par exemple, seules les voyelles moyennes antérieures non-arrondies /e, ɛ/ sont sujettes à l'HV (« l'harmonisation vocalique ne joue pas dans le cas de o inaccentué » p.77). Chez Tranel (1987), il est noté que les voyelles moyennes arrondies /œ/ et /ɔ/ sont moins sensible à l'HV que la voyelle antérieure /ɛ/. Pour /œ/ et /ɔ/, l'HV n'opèrerait que lorsqu'elles sont respectivement suivies de leur contrepartie mi-fermée (comme dans *peureux* [pøRø] ou *auto* [oto]). Cette restriction n'est pas intégrée dans d'autres descriptions, comme celle de Walker (2001) par exemple. S'il relève que l'HV affecte principalement les voyelles /e, ɛ/, elle lui permet d'expliquer aussi l'alternance au sein des paires /o, ɔ/ et /ø, œ/ comme dans *œuvre/œuvrer* [œvR]/[øvRe], *pleut /pleuvoir* [plø]/[plœvwaR] ou *code/coder* [kɔd]/kode], *gros/grossesse* [gRo]/[gRɔsɛs].

Cet effet partagé par les voyelles moyennes avant et arrière ressort également dans les données acoustiques de Nguyen & Fagyal (2008) où les F1 des paires /e, ɛ/ et /o, ɔ/ (les /ø, œ/ n'étant pas examinées) sont abaissés par une voyelle finale fermée ou mi-fermée. Il est à noter que les descriptions diffèrent également en ce qui concerne la directionnalité des modifications subies par la cible : pour certains l'HV en français est uniquement un processus 'fermant' par lequel les voyelles

cibles mi-ouvertes se ferment sous l'influence de la voyelle suivante (Fouché 1959, Tranel 1987, Casagrande 1984), alors que pour d'autres, le processus couvre aussi bien l'influence fermante que peuvent subir les mi-ouvertes que l'influence ouvrante subie par les mi-fermées (p.ex. Malmberg 1969). En lien avec cette définition de la cible potentielle de l'HV, certaines descriptions intègrent des facteurs pouvant interagir avec l'application de l'HV, comme par exemple la loi de position ou des contraintes de fidélité ou d'analogie avec la prononciation de la racine dans les dérivés morphologiques (*coder* [kɔde] < *code* [kɔd]). A celles-ci s'ajoutent des critères orthographiques favorisant telle ou telle prononciation : par exemple, les graphies *é* ou *au/eau* favoriseraient une prononciation mi-fermée ([e] et [o] respectivement) indépendamment de l'aperture de la voyelle suivante (p.ex. Tranel 1989).

La seconde incertitude concerne **la nature de la source (V_2)**, c'est à dire la voyelle finale provoquant l'harmonie. Chez Fouché (1959), seules les voyelles /i, e, y/ ont le pouvoir de fermer la voyelle précédente (seulement /ɛ/ pour lui, ex. *aigre* /ɛgR/ - *aigri* /egRi/). Chez Tranel (1989) la source est définie comme fermée ou mi-fermée mais les exemples qu'il donne n'incluent que des /i, e/ et /ø, o/ (ces dernières influençant que les cibles /œ, ɔ/). Dell (1985), sans préciser un inventaire de V_2, propose des exemples de type *céder* /sede/ - *cédant* /sɛdɑ~/, et semble inclure toutes les voyelles en fonction de leur spécification [+/-bas]. Les résultats de Nguyen & Fagyal (2008) montrent effectivement que les voyelles V_2 moyenne et non-moyenne ont le même effet : l'abaissement du F1 de V_1 est provoqué aussi bien par une V_2 fermée /i/ qu'une mi-fermée /e, ø, o/, et l'augmentation du F1 peut être provoquée aussi bien par une V_2 ouverte /a/ que par /ɛ, œ, ɔ/. Toutefois, ils notent que dans leur corpus, les V_2 non-moyennes (/i-a/) ont un effet systématique sur la fréquence du F1 de V_1 qui est partagé par les six locuteurs, que les V_2 /ø, œ/ ont un effet régulier notamment sur les V_1 postérieures (*porteur/porteuse*), mais que les V_2 /e- ɛ/ n'ont un effet que pour 3 des 6 locuteurs.

Une autre incertitude dans le conditionnement de l'HV concerne la **séquence intervenant entre la cible V_1 et la source V_2**. La question de l'étendue des processus d'harmonie dans les langues a soulevé de nombreux débats dans la littérature, en relation avec la question de savoir si l'HV cible d'une manière séquentielle des segments linéairement ordonnés ou si elle se propage non-linéairement sur des segments partageant la même tier (p.ex. Gafos 1996). Dans la littérature phonétique, la question de l'empan coarticulatoire de voyelle à voyelle a été aussi largement débattue (p.ex. Fowler et Brancazio 2000). Pour l'HV en français, les descriptions sur ce point ne sont pas non plus uniformes. Pour Dell (1985), l'HV nécessite une frontière morphologique entre V_1 et V_2, alors que Nguyen et Faygal (2008) montrent que l'HV opère au travers de celles-ci. Fouché (1959) note que la suite 'rr', un groupe consonantique ou la présence d'un schwa sous-jacent (prononcé ou non) entre V_1 et V_2 bloque l'HV. Dell (1973) au contraire propose une règle d'HV tardive qui s'appliquerait après la délétion du schwa et permettrait donc l'HV (p.ex. *aiderez* [edRe]).

Les nombreuses descriptions que nous avons étudiées convergent toutefois sur un point : l'HV est décrite comme un phénomène optionnel dont l'occurrence est **liée au style de parole**. Elle aurait tendance à être plus fréquente en parole conversationnelle (Fouché 1959), spontanée (Tranel 1987) et peu formelle (Walker 2001). Pour autant, cet aspect n'a jamais été vérifié empiriquement à notre connaissance. Nous examinerons donc le processus d'HV dans deux corpus incluant des styles de parole différents : journalistique, que l'on suppose plus soutenu, et conversationnel, plus relâché.

2 Méthodologie

2.1 Corpus et prétraitements

Deux corpus de grande taille correspondant à différents styles de parole en français standard ont été analysés. Le corpus ESTER (Galliano et al., 2009) comporte des interventions publiques partiellement pré-écrites, extraites de diverses émissions de radio et de TV (bulletins d'information et débats politiques ou de société). La majeure partie de ces données est produite par des locuteurs professionnels avec un degré de contrôle relativement important. Le corpus NCCFr (Torreira et al., 2010) est composé de conversations entre amis sur des sujets de société produites dans une interaction face-à-face informelle. Ces deux corpus ont été alignés automatiquement en mots et phonèmes à partir de transcriptions orthographiques ainsi que segmentés en locuteurs.

A partir d'une liste de mots produits dans ces corpus, nous avons identifié les mots présentant un contexte d'HV potentiel, c'est à dire ayant une $V_2 \in \{i, e, \varepsilon, o, a, y, u, \tilde{a}\}$ en syllabe finale et en syllabe penult une $V_1 \in \{e, \varepsilon, o, ɔ\}$. Aucune contrainte sur la structure syllabique du mot, sa longueur ou sur la séquence de consonnes séparant V_1 et V_2 n'a été apportée. Ainsi, des formes comme *aiderez* alignées sans schwa [V_1dRV_2] sont incluses dans l'analyse. 33 325 mots (corpus NCCFr : 14k mots, ESTER : 19k mots) ont ainsi été sélectionnés.

Les voyelles moyennes V_1 ont été classifiées selon leur lieu d'articulation (et arrondissement, variable *AntérioritéV_1*) avec /e, ɛ/ antérieures et /o, ɔ/ postérieures. Ces voyelles ont été également codées (variable *OrthoV_1*) selon que leur graphie favorise ou non une prononciation mi-fermée : graphie *é* pour [e] et *au/eau* pour [o]. Les voyelles finales V_2 ont été catégorisées en deux degrés d'aperture (variable *ApertureV_2*), [+haut] (fermée ou mi-fermée /i, y, e, u, o/) et [-haut] (mi-ouverte ou ouverte /ɛ, a, ɑ̃/). Etant donné la variabilité des prononciations des V_2/e/-/ɛ/, nous sommes partis des étiquettes phonétiques attribuées par l'alignement forcé (sensés refléter la prononciation mais dépendants des variantes de prononciation prises en compte par le système) que nous avons harmonisées en combinant des informations orthographiques (V_2/e/ pour les graphies *é/és/ée/ées/ez/er* ; V_2/ɛ/ pour les graphies *è/ais/ait/aient* et les quelques cas de substantif en *er* comme laser ou amer) et la transcription phonologique de référence extraite de la base Lexique 3 (New et al., 2007). Les catégories vocaliques sont donc des classes phonémiques indépendantes de leur prononciation dans les corpus. Les voyelles /ø, œ/, dont les relations graphie-phonie sont plus complexes et qui peuvent être confondues avec un schwa, ont été exclues pour cette première étude.

Concernant la séquence de consonne présente entre V_1 et V_2, nous avons distingué plusieurs cas présentant suffisamment d'occurrences dans les différentes catégories définies plus haut. Dans les cas où V_1 et V_2 sont séparées par une seule consonne, nous avons codé (variable *TypeC*) si cette consonne était labiale (*économie*) ou linguale (*échappe*, *écho*). Dans les cas où V_1 et V_2 sont séparées par plusieurs consonnes, nous avons codé dans la variable *SchwaIntervocalique* si la suite de consonne était lexicale, ex. *maîtrise*, *esprit* ou si elle résultait de la chute d'un schwa (c'est à dire les cas où le système d'alignement considère que le schwa n'est pas produit), ex. *laisserez* [lesRe], *souhaiterez* [swetRe].

2.2 Analyse statistique

Les relations entre les valeurs de F1 en Hertz et les effets fixes listés en section 3 ont été testées au moyen de modèles linéaires mixtes (bibliothèque 'lme4' de R, Bates et al., 2014). L' intercept pour

les locuteurs a également été intégré au modèle comme effet aléatoire. De plus, pour éviter un taux élevé d'erreur de type I, les pentes aléatoires par locuteur ont été incluses pour chaque effet fixe, correspondant à la variabilité inter-locuteur de l'effet de chaque facteur fixe sur F1. Les effets aléatoires incluent les pentes pour tous les effets principaux, mais pas pour leurs interactions car cela ne permet pas au modèle de converger. Les valeurs de p (tests de rapport de vraisemblance) sont obtenues par des approximations de type *Satterthwaite* à l'aide de la fonction 'lmerTest', qui permet d'obtenir des estimations plus fiables que les méthodes équivalentes pour les régressions linéaires. 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'.

3 Résultats et discussion

Dans un premier temps, nous avons modélisé les valeurs de F1 en fonction des variables *Aperture*V_2 ([+/-haut]), *Antériorité*V_1 (antérieure/postérieure), *Ortho*V_1 (oui/non) et *Corpus* (journalistique/conversationnel), selon les principes généraux de construction du modèle présentés ci-dessus. L'analyse indique un R^2 marginal de 0.09 et un R^2 conditionnel de 0.36, et met en évidence un effet significatif de chacun des quatre prédicteurs testés ($p<0.0001$ dans les quatre cas). Ainsi, le F1 de V_1 est significativement plus élevé lorsque V_2 est une voyelle basse, lorsque V_1 appartient à la catégorie des voyelles postérieures (/o, ɔ/), lorsque la prononciation de V_1 n'est pas influencée par des critères graphiques favorisant une prononciation mi-fermée, et lorsque le mot est produit en parole journalistique comparativement à la parole conversationnelle. Afin de déterminer quels facteurs interviennent dans l'HV, nous sommes tout particulièrement intéressés aux prédicteurs interagissant avec le facteur *Aperture*V_2. L'HV étant mesurée comme l'abaissement du F1 de V_1 induit par la présence d'une V_2 /+haut/ (fermée ou mi-fermée) par rapport à une V_2 /-haut/ (ouverte ou mi-ouverte), ces interactions nous permettent de comparer le degré d'HV associé aux différentes valeurs de ces prédicteurs.

Concernant le lieu d'articulation de V_1, nous observons une interaction significative entre *Antériorité*V_1 et *Aperture*V_2 ($p<0.0001$). Ces interactions, ainsi que la quantification de l'HV pour les deux types de V_1 considérés, sont illustrées par la Figure 1. Le degré d'HV, i.e. l'effet de V_2 sur le F1 de V_1, est plus important pour les voyelles postérieures arrondies /o, ɔ/ que pour les voyelles antérieures /e, ɛ/. Ce résultat est en désaccord avec les descriptions classiques présentées dans la littérature (Fouché, 1959 ; Tranel, 1987). D'une part, nous observons que le timbre des voyelles postérieures moyennes est aussi affecté par l'aperture de la voyelle suivante, contra Fouché (1959) mais conformément aux résultats empiriques de Nguyen et Fagyal (2008). D'autre part, contra Tranel (1987) nous montrons que celles-ci sont plus sensibles à l'HV, avec une HV supérieure pour les voyelles moyennes postérieures vs. antérieures. Nguyen et Fagyal (2008) n'ont pas fait de comparaison directe entre V_1 postérieure et antérieure, mais si pour les deux ils observent un effet de V2 sur V_1, ils relèvent un effet différent sur F2 : le F2 de /e-ɛ/ augmente alors que celui de /o-ɔ/ s'abaisse sous l'influence d'une V2 [+haute].

Notre analyse met également en évidence une interaction significative entre *Aperture*V_2 et *Ortho*V_1 ($p<0.0001$), montrant un degré d'HV plus important dans les formes où la graphie de V_1 ne l'empêche pas d'opérer, c'est à dire quand la graphie favorise une prononciation mi-fermée. On remarque que les graphies *é/au/eau* favorisent effectivement une prononciation plus fermée de V_1, avec un F1 plus bas, indépendamment de la voyelle suivante.

Ce conditionnement orthographique de la prononciation de V_1 ressort également de façon surprenante dans une interaction significative entre *Ortho*V_1 et *Corpus* ($p<0.0001$) montrant que la

différence de F1 entre une V_1 marquée orthographiquement ou non est plus grande en parole conversationnelle par rapport à la parole journalistique. Ceci va à l'encontre de l'hypothèse selon lesquels les modulations des représentations phonologiques induites par l'orthographe sont plus fortes dans un style de parole plus formel (Taft & Hambly, 1985). Si on considère que les locuteurs professionnels du corpus ESTER lisent en partie des contenus scriptés, on pourrait s'attendre à une plus forte influence orthographique dans ce corpus et chez ces locuteurs ayant probablement une plus grande maîtrise de l'écrit que les jeunes étudiants du corpus NCCFr (cf. Olson 1996).

Concernant l'effet du style de parole sur l'HV, ici encore les résultats sont surprenants. Une interaction significative entre *Aperture*V_2 et *Corpus* ($p<0.0001$) fait ressortir que l'effet de V_2 sur le F1 de V_1 est globalement plus important en parole journalistique qu'en parole conversationnelle. Ceci va à l'encontre des descriptions antérieures pour lesquelles l'HV serait plus marquée en parole peu formelle. Toutefois, nous remarquons également que ces deux facteurs sont impliqués dans une interaction à trois niveaux avec le facteur *Ortho*V_1 (*Aperture*V_2**Ortho*V_1**Corpus*, $p<0.0001$). Comme illustré Figure 2, l'effet du corpus sur l'HV interagit avec les contraintes orthographiques : si l'on ne considère que les voyelles V_1 non marquées orthographiquement, l'HV ressort comme plus importante dans le corpus de parole conversationnelle par rapport au corpus journalistique.

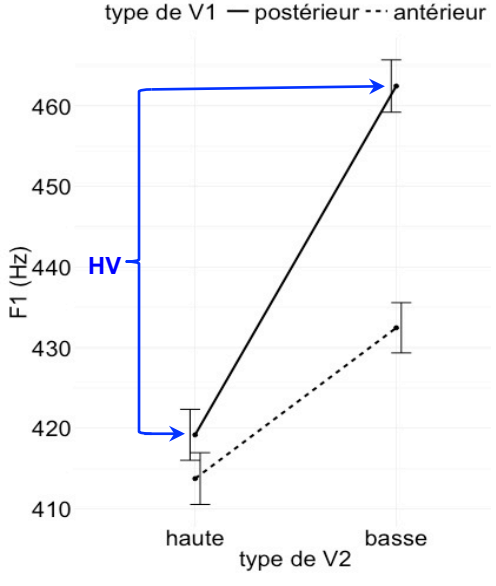


FIGURE 1 : F1 de V_1 (ajustées par le modèle, moyenne et erreur type) en fonction de l'aperture de V_2 et de l'antériorité de V_1. En bleu, la différence de F1 entre un contexte V_2 haut vs. bas illustre la mesure du degré d'HV pour les V_1 postérieures.

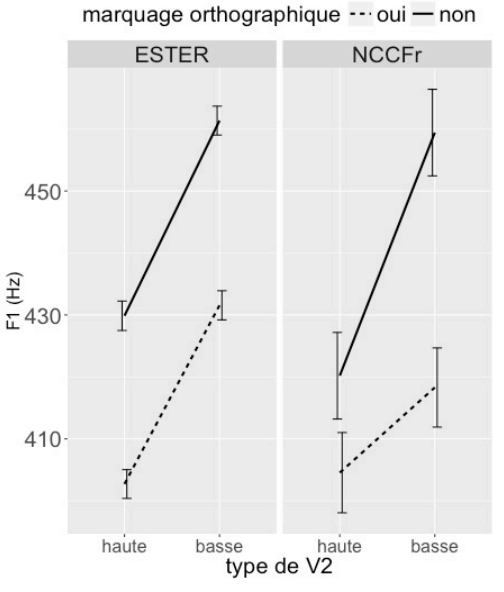


FIGURE 2 : F1 de V_1 (ajustées par le modèle, moyenne et erreur type) en fonction de l'aperture de V_2, du corpus et des interférences orthographiques.

Afin de tester des hypothèses plus spécifiques sur les contextes susceptibles de favoriser ou non l'HV, nous avons construit deux modèles supplémentaires pour évaluer les liens entre l'HV et les éléments intervocaliques : à savoir (a) si l'articulation de la consonne, lorsqu'elle est seule, laisse la langue libre d'anticiper les mouvements linguaux entre les voyelles (consonne labiale) par rapport à une consonne nécessitant une articulation linguale, et (b) s'il existe entre les deux voyelles un autre

élément vocalique (un schwa) sous-jacent même s'il n'est pas prononcé. La variation relative à l'antériorité/postériorité de V_1 a également été prise en considération dans ces modèles en incluant *AntérioritéV1* comme facteur aléatoire, seules les voyelles V_1 non marquées orthographiquement étant incluses dans l'analyse. Notons toutefois que pour ces deux modèles, nous obtenons des résultats très similaires en termes d'effets des facteurs et interactions pour les voyelles V_1 marquées orthographiquement comme mi-fermées.

Nous avons tout d'abord testé l'influence du type de consonne intervocalique sur l'HV lorsque V_1 et V_2 ne sont séparés que par une consonne. Pour cela, nous avons modélisé les valeurs de F1 en fonction des variables *ApertureV2* et *TypeC* (labiale ou linguale). L'analyse indique un R^2 marginal de 0.02 et un R^2 conditionnel de 0.38, et ne montre pas d'effet significatif d'*ApertureV2* (p=0.18) ni de *TypeC* (p=0.63). En revanche, elle montre une interaction significative entre les deux facteurs (p<0.0001). Comme le montre la Figure 3, l'HV est plus importante quand V_1 et V_2 sont séparées par une consonne labiale plutôt qu'une linguale : la différence de F1 entre V_2 [-haute] et [+haute] est plus élevée dans le cas d'une consonne labiale. Ce résultat confirme sur de plus grands corpus les travaux bien connus sur les liens entre coarticulation voyelle-à-voyelle et résistance à la coarticulation (cf. par exemple Recasens, 1985 ou encore Fowler & Brancazio, 2000). Ces études montrent que la coarticulation voyelle-à-voyelle est favorisée au travers d'une consonne 'faiblement résistante' à la coarticulation telle qu'une labiale qui n'implique pas de constriction linguale.

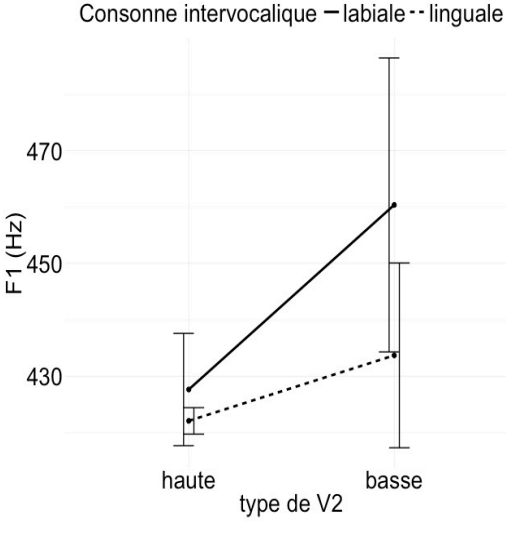


FIGURE 3 : F1 de V_1 (ajustées par le modèle, moyenne et erreur type) en fonction l'aperture de V_2 et du type de consonne intervocalique (labiale vs. linguale).

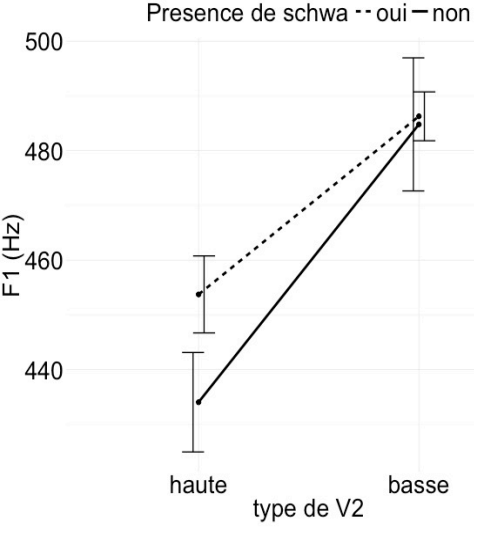


FIGURE 4 : F1 de V_1 (ajustées par le modèle, moyenne et erreur type) en fonction l'aperture de V_2 et de la présence ou non d'un schwa sous-jacent.

Enfin, notre dernière analyse permet de clarifier le rôle controversé d'un schwa intervocalique non-prononcé dans l'HV, en modélisant F1 en fonction des variables *ApertureV2* et *SchwaIntervocalique*. Pour cela, nous n'avons pris en considération que les cas dans lesquels V_1 et V_2 sont séparées par deux consonnes ou plus, soit les cas exclus dans le modèle précédent. L'analyse indique un R^2 marginal de 0.06 et un R^2 conditionnel de 0.34, et montre un effet significatif d'*ApertureV2* (p<0.0001) mais pas d'effet significatif de *SchwaIntervocalique* (p=0.15).

Elle montre également une interaction significative entre *ApertureV_2* et *SchwaIntervocalique* ($p<0.05$), illustrée par la Figure 4. On remarque que les modulations de F1 en fonction de V_2 sont plus importantes quand les deux voyelles ne sont pas séparées par un schwa sous-jacent. On remarque également que quand V_2 est [-haute], le timbre de V_1 semble mi-ouvert dans les deux conditions et que ce n'est quand V_2 est [+haute] que l'abaissement du F1 de V_1 semble freiné par la présence d'un schwa sous-jacent.

Plusieurs interprétations de ce résultat sont possibles. La première serait que la catégorie 'schwa présent' comprenne des productions dans lesquelles le schwa a vraiment été prononcé sans être détecté par l'alignement automatique, et donc ne seraient pas des contextes d'HV. La seconde serait que la présence d'une cible vocalique sous-jacente, même non réalisée, interfère dans cette interaction entre V_1 et V_2, en étant opaque à l'harmonie. Pourtant, dans nos données l'HV, même si elle est réduite, ne semble pas absente dans les cas avec schwa sous-jacent (contra Fouché, 1959). Un examen plus détaillé des mots inclus dans cette catégorie sera nécessaire pour mieux comprendre cette interaction.

4 Conclusion

L'objectif de cette étude était de mieux décrire les contextes segmentaux, orthographiques et les effets de style, susceptibles d'après la littérature d'interagir avec l'HV en français. Les variations de timbre (déterminée par leur F1) des voyelles moyennes {e, ɛ, o, ɔ} ont été examinées en fonction de l'aperture de la voyelle finale suivante, ouverte/mi-ouverte vs. fermée/mi-fermée dans deux gros corpus de parole journalistique et conversationnelle. Le F1 des voyelles cibles varie en fonction des voyelles suivantes (plus bas devant V_2 [+haute] et plus haut devant V_2 [-haute]) et donc l'HV paraît clairement opérer dans nos données. Cet effet est modulé par le lieu d'articulation de V_1 (plus fort pour /o-ɔ/) et peut être atténué par des interférences orthographiques et la séquence intervenant entre la source et la cible.

Nos prochains travaux testeront également dans quelle mesure le degré d'HV change en fonction de la proéminence prosodique, et plus spécifiquement si les voyelles cibles situées en position prosodique forte sont plus résistantes à l'HV. Ceci nous permettra d'approfondir notre compréhension de la relation entre coarticulation voyelle-à-voyelle et HV, et à un niveau plus général sur les liens entre manifestations phonétiques et processus phonologiques.

Enfin, au-delà de la nécessité déjà soulignée par Nguyen et al. (2004) d'établir la pertinence de l'HV en français sur le plan perceptif, les études que nous prévoyons de mener sur cette thématique exploreront la capacité d'auditeurs francophones à exploiter la coarticulation vocalique anticipatoire pour améliorer les performances de l'accès lexical (Tobin et al., 2010).

Remerciements

Ce travail a été financé par le programme de recherche et d'innovation de l'Union Européenne Horizon 2020, à travers la bourse Marie Sklodowska-Curie n°662530 accordée au premier auteur et par le programme "Investissements d'Avenir" géré par l'Agence Nationale de la Recherche ANR-10-LABX-0083 (Labex EFL).

Références

BATES, D. MAECHLER, M., BOLKER, B. and WALKER, S. (2014). lme4: Linear mixed-effects models using eigen and s4 . R package version 1.1-7. http://CRAN.Rproject. org/package=lme4.

DELL, F. (1972). Les règles et les sons: Introduction à la phonologie générative. Paris: Hermann

FAGYAL, Z., NGUYEN, N., BOULA DE MAREÜIL, P. (2003). From dilation to coarticulation: is there vowel harmony in French? *Studies in Linguistic Sciences*, 32, 1-21.

FOUCHÉ, P. (1959). Traité de prononciation française. Paris: Klincksieck

FOWLER C. A. et BRANCAZIO, L. (2000) Coarticulation resistance of American English consonants and its effects on transconsonantal vowel-to-vowel coarticulation. *Language and Speech*, 43(1), 1-41.

GAFOS, I. A. (1996). The articulatory basis of locality in phonology. (Doctoral dissertation), Johns Hopkins University [Published 1999, New York, NY: Garland].

GALLIANO, S., GRAVIER, G., CHAUBARD, L. 2009. The ESTER 2 evaluation campaign for the rich transcription of French broadcasts. *Proceedings of Interspeech*, Brighton (UK), 2583–2586.

GRAMMONT, M. (1926). La prononciation française. Delagrave.

MALMBERG B. (1969) Phonétique Française, Malmô, Hermods

NGUYEN, N., FAGYAL, Z., and COLE, J. (2004). Perceptual relevance of long-domain phonetic dependencies. *Proceedings of the IVth Linguistic Studies* Workshop, 173–178, Nantes, France.

NGUYEN, N., FAGYAL, Z. (2008). Acoustic aspects of vowel harmony in French. *Journal of Phonetics*, 36, 1-27.

NEW, M. BRYSBAERT M., VERONIS J., PALLIER C. (2007) *The use of film subtitles to estimate word frequencies*, Applied psycholinguistics, 28(04), 661–677.

OHALA, J. (1994). Towards a universal, phonetically-based theory of vowel harmony. *Proceedings of the ICSLP,* Yokohama. 10-14.11.-10-14.14

OLSON, D. R. (1996) Towards a psychology of literacy: On the relations between speech and writing. *Cognition*, 60, 83–104.

RECASENS, D. (1985). Coarticulatory patterns and degree of coarticulation resistance in Catalan CV sequences. *Language and Speech*, 28, 97-114.

TAFT, M., HAMBLY, G. (1985). The influence of orthography on phonological representations in the lexicon. *Journal of Memory and Language*, 24(3), 320-335.

TILSEN, S. (2007). Vowel-to-vowel coarticulation and dissimilation in phonemic-response priming. UK Berkeley Phonology Lab Annual report. 416-458

TOBIN, S. J., CHO, P. W., JENNETT, P. M., Magnuson, J. S. (2010). Effects of anticipatory coarticulation on lexical access. *Proc. Mtgs. Cognitive Sci,* 2200-2205.

TRANEL, B. (1987). The Sounds of French: An Introduction. Cambridge: CUP.

TORREIRA F., ADDA-DECKER M., ERNESTUS, M. (2010). The Nijmegen Corpus of Casual French. *Speech Communication*, 52, 201-221.

WALKER, D. C. (2001). French Sound Structure. Calgary: University of Calgary Press.

Quelle(s) mesure(s) de similarité prosodique comme évaluation de l'imitation ?

Olivier Nocaudie Corine Astésano

U.R.I. Octogone-Lordat (E.A. 4156), Université de Toulouse, UTM

nocaudie@univ-tlse2.fr, corine.astesano@univ-tlse2.fr

RESUME

La performance imitative des locuteurs varie de celle du professionnel, expert, à celle du naïf, plus ou moins talentueux. L'étude de l'imitation souligne la difficulté pour trouver des indices mesurables de la réussite d'une imitation. Dans cette étude exploratoire, des contours de *f0* recueillis au fil de tâches d'imitation sont testés au moyen d'une double approche : mesure objective par le biais de deux mesures de la similarité prosodique reportées dans la littérature et évaluation perceptive par un panel de 15 auditeurs naïfs. Nos premiers résultats indiquent une bonne corrélation entre les deux approches et soulèvent la question du choix de l'indice mesurable qui rendrait le mieux compte d'une imitation au niveau tonal. Ils soulignent également la variabilité interindividuelle des comportements imitatifs en parole tout en ouvrant des perspectives intéressantes dans le domaine de la formation à la phonétique corrective par la Méthode Verbo-tonale.

ABSTRACT

Which measure(s) of prosodic similarity as an evaluation of imitation?

Imitative proficiency across speakers is highly variable. Studies on imitation underline how difficult it is to find measurable cues to assess a successful imitation. In this exploratory study, *f0* contours collected from imitations tasks are tested in a double approach: objective measurements of prosodic similarity using two measures reported in the literature and perceptive evaluation by a panel of 15 naïve listeners. Our first results indicate a good correlation between the two approaches and they raise the question concerning the selection of the measurable factor assessing a successful imitation at a tonal level. Furthermore, these results underline an imitative proficiency's variability across speakers while opening perspectives in the domain of phonetic correction using the Verbo Tonal Method.

MOTS-CLES : parole, prosodie, imitation, mesures objectives, évaluations perceptives.

KEYWORDS: speech, prosody, imitation, objective measurements, perceptive evaluation.

1 Introduction

Les études sur l'imitation de la parole rapportent différents types de comportements imitatifs, comme la convergence (une adaptation mutuelle des interlocuteurs au fil de la conversation, Pardo,

2006), l'impersonnation (la tentative d'usurper la voix de l'autre, Révis, De Looze, & Giovanni, 2013), la simple imitation (Mixdorff, Cole, & Shattuck-Hufnagel, 2012) ou le *shadowing* (Dufour & Nguyen, 2013; Goldinger, 1998). Définir ces comportements en se basant sur des facteurs comme le contexte de production (Lewandowski, 2012) ou l'intention de l'imitateur (Donald, 1993) aboutit à les différencier. Pourtant, ils partagent une similitude définitoire majeure : pour être qualifiés de comportements imitatifs, la production du locuteur doit être perçue par un tiers comme similaire au modèle de l'imitation. Par conséquent, l'étude de l'imitation en parole vise à faire produire, puis à observer des changements comportementaux dans la manière de parler de locuteurs (naïfs ou experts), au niveau lexical ou phonétique. Pour ce qui est du niveau phonétique, deux questions ne cessent de représenter un défi méthodologique : quels sont les paramètres du signal sonore perçus puis imités en priorité par les locuteurs ; de quelle manière évaluer et comparer les paramètres choisis, entre leur modèle et leur(s) imitation(s) ? Choisir quel(s) paramètre(s) acoustique(s) mesurer et lier aux résultats d'évaluations perceptives de l'imitation demeure un choix crucial (Pardo, 2013).

Les imitateurs professionnels parviennent à ajuster globalement leur voix aux spécificités de leurs voix cibles, mais ils sont aussi capables d'imiter en reproduisant des variations instantanées de contours intonatifs ou de fréquence et durée des pauses (stratégies de synchronie) (Révis et al., 2013). En revanche, d'après ces derniers travaux, les locuteurs naïfs semblent limités à des stratégies d'ajustement global. Cette dernière remarque soulève une série de questions connexes, en lien avec les stratégies de synchronie : **(1)** jusqu'à quel point un locuteur naïf est-il capable de reproduire un patron prosodique perçu (variation instantanée) ; **(2)** est-il possible d'entraîner un locuteur à reproduire fidèlement des contours intonatifs ; **(3)** comment peut-on évaluer leur réussite –et leur échec– dans l'accomplissement de cette tâche

Les deux premières questions ont une pertinence certaine dans le domaine de l'enseignement de la prononciation à des locuteurs de langue seconde (L2), plus particulièrement en se plaçant dans le cadre théorique proposé par la Méthode Verbo-Tonale d'intégration phonétique (MVT). La MVT postule que les erreurs de prononciation en L2 seraient dues à un biais de perception de la L2. Afin de neutraliser les effets de ce biais, la MVT propose une rééducation perceptive, au moyen d'un ensemble de procédés correctifs où l'influence de la prosodie sur les segments phonétiques joue un rôle majeur. Un enseignant recourant à la MVT doit donc avoir une conscience et un contrôle prosodique efficace, notamment lorsqu'il doit produire des énoncés délexicalisés afin de faciliter la perception par l'apprenant des caractéristiques rythmiques et intonationnelles de la langue cible (Billières, Alazard, Astésano, & Nocaudie, 2013).

Intrinsèquement, une séquence de correction phonétique représente un cas typique d'interaction imitative. En effet, durant une interaction MVT, l'enseignant comme l'apprenant doivent imiter ou réitérer des sons de leur interlocuteur. L'apprenant doit répéter le modèle proposé par l'enseignant, ce qui peut conduire à questionner le lien entre perception et (re)production de la parole chez le sujet devenant bilingue. De son côté, l'enseignant doit produire systématiquement des patrons prosodiques phonologiquement cohérents, soulevant ainsi la question du contrôle de sa production, plus particulièrement au niveau mélodique.

Si les questions **(1)** et **(2)** peuvent être liées à la fois à l'enseignant et à l'apprenant, la présente étude se focalise sur la capacité de l'enseignant à reproduire systématiquement des patrons prosodiques. En effet, avant même de pouvoir tester la capacité de l'apprenant à (re)produire les paramètres phonétiques d'un énoncé, il faut s'assurer que l'enseignant même est capable d'imiter les éléments prosodiques saillants de la parole. Or, la pratique de la MVT implique la reproduction de paramètres

prosodiques de manière maîtrisée et la mise en valeur de certains événements pour faciliter à l'apprenant la perception des sons cibles. Ainsi, cette étude propose de tester en premier lieu la capacité des locuteurs de L1 à produire des imitations de contours prosodiques. Ce faisant, nous nous intéresserons plus particulièrement à notre question **(3)**, à savoir : évaluer le degré de réussite d'une imitation. La méthode d'évaluation de la (dis)similarité prosodique pourrait conduire à la création d'un outil d'évaluation du contrôle prosodique de l'enseignant de L2, et ainsi être utile à leur formation dans le domaine de la correction phonétique.

La littérature sur l'imitation parolière en langue française est réduite, et peu d'études ont abordé plus spécifiquement la reproduction des indices prosodiques (voir cependant Michelas & Nguyen, 2011 à propos de l'accent initial). Cette communication est la poursuite d'une autre étude préliminaire qui décrivait la capacité de locuteurs à imiter les indices prosodiques de phrases contrôlées, au cours de trois tâches d'imitation, d'une simple répétition à une exagération (Nocaudie & Astésano, 2012).

2 Matériel linguistique : un corpus d'imitation(s)

Notre corpus d'imitation est issu d'un corpus de phrases (*Corpus d'Edimbourg : CE*) présentant une ambigüité syntaxique qui peut être résolue par la production des indices prosodiques pertinents. L'ambigüité syntaxique dérive de la manipulation de la portée de l'adjectif, comme dans « les gants et les bas lisses », où l'adjectif (A) « lisses » qualifie alternativement le second nom « bas » uniquement ([les gants][et les bas lisses]) (*Condition 1 : Cond-1*) ou bien les deux noms ([les gants et les bas][lisses]) (*Condition 2 : Cond-2*). La longueur des noms et des adjectifs varie de une à quatre syllabes. Le contrôle de l'ambigüité syntaxique et de la longueur des constituants nous permet d'observer les indices prosodiques (proéminences, frontières, tons, pauses…) utilisés dans la linéarisation syntaxique des énoncés oraux (voir Astésano, Bard, & Turk, (2007) pour le détail de la constitution du *CE)*.

Du *CE*, 16 énoncés prononcés par une locutrice ont été sélectionnés comme stimuli auditifs pour la constitution du corpus *Imitation* (*CI*). Ces phrases comportaient uniquement deux longueurs de noms (noms tri- et quadrisyllabiques) combinées à des adjectifs de une à quatre syllabes, prononcées dans les deux conditions syntaxiques (*Cond-1* & *Cond-2*). 8 locuteurs naïfs ont dit ces phrases au fil de trois tâches différentes : a) une simple répétition (*REP*), b) une imitation (*IMI*), c) une exagération des phrases de la locutrice (*EXA*). Les données de 2 locuteurs ont été exclues du *CI* en raisons de facteurs physiologiques ayant eu une incidence sur leur voix (stress induit par la situation expérimentale, timbre éraillé). Durant chaque tâche, les imitateurs répétaient chaque phrase 3 fois, dans un ordre aléatoire, nous permettant d'obtenir un total de 864 phrases (16 énoncés * 2 conditions syntaxiques * 3 occurrences * 3 tâches * 6 imitateurs). Afin d'évaluer la capacité implicite des locuteurs à imiter la parole, l'attention des locuteurs n'était pas focalisée sur le fait d'imiter durant la tâche a). Il leur était simplement demandé de « *dire la phrase entendue en préservant sa structure* ». Durant les tâches b) & c), il leur était demandé explicitement de produire une imitation ou une exagération des énoncés entendus. Cependant, l'expérimentateur n'a pas dirigé l'attention des imitateurs sur les indices prosodiques.

La présente étude vise à comparer des données perceptives (Test AX) et objectives (issues d'un algorithme) obtenues à partir d'un extrait du *CI*. Nous avons sélectionné un sous-corpus de 4 énoncés de 2 longueurs différentes pour tester la robustesse de notre algorithme de mesure de la similarité prosodique. En effet, toutes les phrases sont prises dans la condition *Cond-1* car la désambigüisation syntaxique est normalement marquée par une pause silencieuse entre le premier et

le second nom. Comme l'algorithme a une nette tendance à aligner les silences lorsqu'il aligne les contours, l'absence de pause dans l'imitation devrait gêner l'alignement avec la référence et se traduire théoriquement par un score de similarité moins élevé.

Les tests ont été menés sur la production de 4 imitateurs (Sp1, Sp3, Sp5 & Sp7), qui étaient appariés pour prononcer certains énoncés. Sp1 (femme) & Sp5 (femme) ont dit « *Les bagatelles et les balivernes sottes* » et « *Les bonimenteurs et les baratineurs fades* » ; Sp3 (homme) & Sp7 (femme) ont prononcé « *Les bagatelles et les balivernes saugrenues* » et « *Les bonimenteurs et les baratineurs fabuleux* ». Nos résultats ont été calculés sur 18 énoncés par locuteur, produisant alors un nombre total de 72 couples de phrases (modèle *vs.* imitation) : [(2 énoncés * 3 répétitions * 3 tâches) * 4 locuteurs].

3 Méthode : mesures objectives & évaluations perceptives d'imitations prosodiques

Un problème constant dans l'évaluation de l'imitation en parole réside dans la relative absence de congruence entre la multitude de paramètres acoustiques pouvant être mesurés, alors que ces derniers divergent ou convergent avec le modèle. La mélodie, et son corrélat physique, la *f0* ont été décrits comme une cible de choix pour l'imitateur (Révis et al., 2013). Par ailleurs, les procédés correctifs de la MVT reposent largement sur la manipulation des contours mélodiques. Ainsi, notre méthode se concentre principalement sur la mesure de la distance physique de contours de *f0* appariés (modèle *vs.* imitation) d'une part et sur l'évaluation perceptive de leur ressemblance d'autre part.

3.1 Dynamic Time Warping (DTW) & (dis)similarité prosodique

Mesurer objectivement l'imitation d'un contour mélodique équivaut à trouver s'il y a une distance physique entre le contour modèle et sa reproduction, *i.e.* connaître le degré de correspondance entre les deux formes de contours intonatifs.

Ceci étant dit, comparer des formes impose une normalisation tonale et un alignement temporel des pics et des creux décrits par les contours de *f0* (DTW). Cette méthode d'interpolation non linéaire, qui force l'alignement entre le modèle et sa reproduction, améliorerait les scores de corrélation entre les formes, notamment si les contours intonatifs sont fonctionnellement similaires, *i.e.* s'ils partagent le même patron accentuel (Rilliard *et al.*, 2011). La distance entre les paires de contours intonatifs a été calculée au moyen de deux mesures similaires à celles proposées par Hermes (1998) où : *w(t)* correspond au décours temporel du facteur de poids (la somme des *subharmonic sumspectrum* du contour imité), *W* est l'intégrale du contour de 0 à *T* (avec *T*, la durée totale du contour) et *f1* & *f2*, la paire de contours testée par l'algorithme.

Pour ces travaux, la procédure de normalisation de la *f0* choisie diffère de celle d'Hermes. En effet, nous avons divisé chaque valeur de *f0* par la valeur maximale de *f0* de l'énoncé (*f1 = 1/p1max*). Cette procédure de normalisation permet de comparer plus facilement les imitateurs masculins et féminins en ramenant la variation des courbes de *f0* dans des valeurs comprises entre 0 et 1. Par ailleurs, cette procédure de normalisation facilite le travail de la partie DTW de l'algorithme. Les valeurs de *f0* ont été extraites à l'aide du logiciel PRAAT (Boersma, 2001) et le taux d'échantillonnage était d'une valeur par milliseconde (Rilliard, Allauzen, & de Mareüil, 2011).

Après normalisation et DTW, nous avons calculé la différence de la moyenne des moindres carrés (ci-après, L_2) de chaque paire de contours comme suit :

$$L_2 = \left\{\frac{1}{W}\int_0^T w(t)\left|f_1(t) - f_2(t)\right|^2 dt\right\} \qquad (1)$$

Le coefficient de corrélation *r* entre deux contours *f1* et *f2* ont ensuite été calculés de la manière suivante :

$$r = \frac{\frac{1}{W}\int_0^T w(t) f_1(t) f_2(t) dt}{\sqrt{\left\{\frac{1}{W}\int_0^T w(t)\left|f_1(t)\right|^2 dt \frac{1}{W}\int_0^T w(t)\left|f_2(t)\right|^2 dt\right\}}} \qquad (2)$$

Afin de pouvoir comparer les coefficients de corrélation, Hermes propose de transformer *r* en *Z* de Fisher (ci-après, *Zr*), que nous calculons ainsi :

$$Z_{f_1 f_2} = \frac{1}{2} ln \frac{1 + r_{f_1 f_2}}{1 - r_{f_1 f_2}} \qquad (3)$$

L_2 mesure les changements rapides de *f0*, tandis que *Zr* est une mesure holistique de la proximité de la forme de deux contours. Il est intéressant de tester leur complémentarité vis-à-vis des jugements perceptifs que nous recueillons par ailleurs.

Finalement, chaque phrase reproduite a reçu un rang en fonction de ses scores *L2* et *Zr*. L_2 est une mesure de dissimilarité (soit, plus L_2 a une valeur élevée, plus grande est la dissimilarité entre le modèle et sa reproduction). La phrase avec le L_2 le plus bas a donc été classée comme rang 1, celle avec le L_2 le plus haut, rang 72. *Zr* est une mesure de similarité (plus *Zr* est haut, plus la similarité entre les deux contours est grande). La phrase avec le plus haut *Zr* a donc reçu le rang 1, celle avec le plus bas, le rang 72. Ainsi, nous avons obtenu 2 classements différents (en fonction de L_2 ou de *Zr*), qui seront comparés au classement dérivé des résultats des évaluations perceptives.

3.2 Test AX de jugement de la similarité

Comme le note Pardo (2013), l'imitation en parole devrait être évaluée objectivement et subjectivement, *i.e.* physiquement et perceptivement. A cette fin, nous avons complété les mesures objectives décrites ci-dessus avec des données issues d'une tâche AX de jugement de la similarité. Cette tâche permet d'obtenir une évaluation holistique de chaque phrase imitée (X) comparée à son modèle (A). 15 auditeurs naïfs ont pris part à la tâche. Tous étaient de L1 française (âge : 25-32) et ne présentaient pas de trouble de la parole ou de l'audition.

Il était demandé aux locuteurs de noter la ressemblance de X avec A en termes de musicalité de la parole (rythme, variation tonale). Le test AX a été passé au moyen du logiciel Lancelot (environnement HTML de PERCEVAL (André, Ghio, Cavé, & Teston, 2003)). Les paires de phrases étaient randomisées par le logiciel, et présentées en modalité auditive dans des écouteurs de qualité professionnelle. Les auditeurs pouvaient écouter chaque couple de phrase jusqu'à cinq fois avant de leur attribuer une note sur une échelle Likert allant de 1 (moins similaire) à 5 (très similaire) en touchant l'écran tactile de l'ordinateur de passation.

Pour chaque phrase X, la moyenne des 15 scores obtenus a été calculée. La phrase obtenant la moyenne la plus élevée a obtenu le rang 1 du classement AX, etc. En cas d'égalité de score entre deux phrases X ou plus, un rang égal à la moyenne des rangs qu'elles devraient occuper dans le classement a été attribué à chacune d'entre elles (par exemple : 7 ; 8 ; 9 → 8 ; 8 ; 8).

4 Résultats

Nous décrirons tout d'abord les résultats comparant les classements obtenus à partir des scores objectifs (L_2 & Zr) et du score perceptif (AX) pour tous les locuteurs. La Figure 1 montre la distribution des rangs pour les trois types de scores, qui donneront par la suite lieu à un calcul de corrélation. Sur ce diagramme, le point indique le score moyen obtenu par le locuteur et les boîtes montrent la distribution interquartile des rangs de chacune des 18 phrases prononcées par locuteur, parmi l'ensemble de 72 phrases évaluées. Les barres de confiance montrent les valeurs minimales et maximales de rangs obtenus par chaque locuteur, en fonction du classement. Nous rappelons que le rang 1 montre la production qui a obtenu le meilleur jugement de similarité.

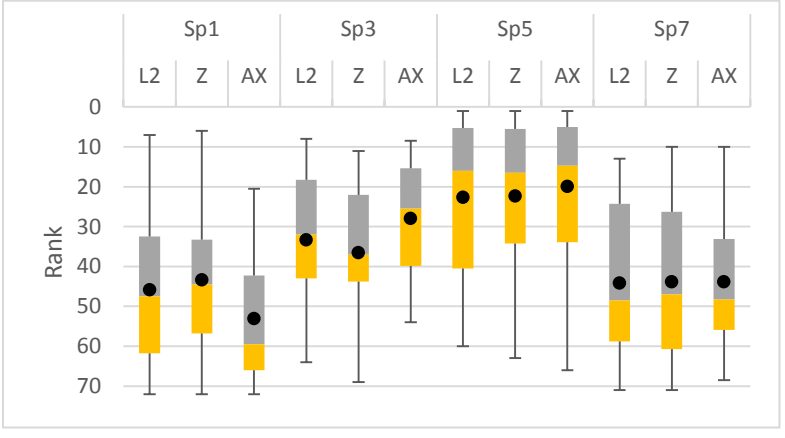


Figure 1 : Distribution des rangs (de 1 à 72, en ordonnées) obtenus par chaque locuteur (Sp1, Sp3 ; Sp5 & Sp7) en fonction des types de scores (L_2, Zr & AX). Les points représentent le rang moyen de chaque locuteur pour les 18 énoncés imités.

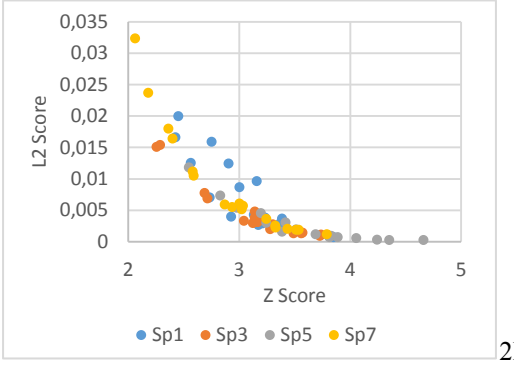
A la vue de cette distribution des rangs obtenus par chaque locuteur, il semble raisonnable de classer Sp5 comme le locuteur le plus performant dans les tâches d'imitation (Rang moyens : L_2 = 22,76 ; Zr = 22,28 ; AX = 19,88), suivi clairement par Sp3 (L_2 = 33,33 ; Zr = 36,56 ; AX = 27,92). Les rangs des classements objectifs de Sp7 (L_2 = 44,17 ; Zr = 43,83) et de Sp1 (L_2 = 45,83 ; Zr = 43,33) sont très proches, mais leur rangs du classement AX (Respectivement : AX(Sp7) = 43,82 ; AX(Sp1 = 53,06) peut refléter la plus grande dispersion de leurs rangs dans le quartile inférieur : le meilleur rang obtenu par Sp1 est meilleur que celui de Sp7, mais cette valeur isolée peut fausser le calcul de la moyenne de leurs rangs. De fait, Sp7 a obtenu un plus grand nombre de bons rangs que Sp1 au fil des procédures d'évaluation, comme le résume le diagramme.

Le calcul de la corrélation a été fait au moyen de Real Statistics pour Excel (Zaiontz, 2015). Afin de comparer les données objectives et subjectives, nous proposons d'utiliser le r_s de Spearman, qui est un indice de corrélation entre des données *ordonnées* (rangs des classements). Les tests bilatéraux de Spearman menés sur nos données montrent des corrélations positives :

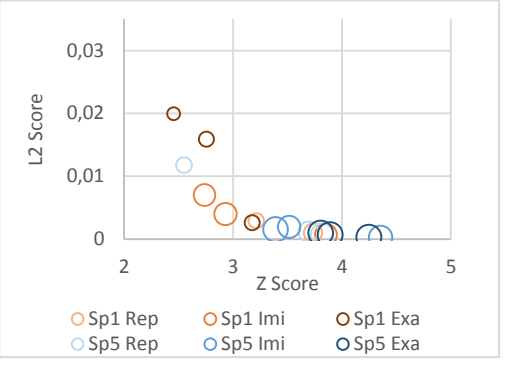
- Zr vs. AX → r_s = .554, p < .0001, t (71) = 5.562
- L_2 vs. AX → r_s = .589, p < .0001, t (71) = 6,092

Les deux indices de corrélation dépassent les valeurs critiques de r_s admises pour N = 72 (*rs-crit* = .382, *t-crit* = 3.43). Ainsi, la relation linéaire entre les classements objectifs et perceptifs semble robuste.

La Figure 2A montre la relation entre L_2 et *Zr* pour les 72 phrases testées. Les points dans le coin en bas à droite sont les phrases qui ont été jugées très similaires à leur modèle. Les résultats donnés par l'algorithme soulignent la différence de performance en imitation entre les différents locuteurs. La Figure 2B illustre la différence de performance entre les deux locuteurs qui ont été respectivement classés le moins (Sp1) et le plus (Sp5) performants au fil des tâches, d'après les résultats de l'algorithme et du panel d'auditeurs.



2A



2B

Figure 2 (A & B) : La Figure 2A montre les scores L_2 (axe *y*) & Zr (axe *x*) de 71 phrases. La Figure 2B illustre en trois dimensions l'évaluation de la performance de 2 locuteurs imitant la phrase « Les bagatelles et les balivernes sottes » au cours de REP, IMI & EXA. Les scores Z_r et L_2 sont respectivement sur les axes des *x* et des *y* (plus L_2 est haut, moins la paire est similaire ; plus Z_r est haut, plus la paire est similaire), la taille des points représente le score moyen au test AX. Un point d'exception a été retiré de ces représentations (Sp1 Rep L2=0.0615, Zr=1.6903, AX =1.4) afin d'améliorer la lisibilité de ces représentations.

Nous prévoyions que les imitations les plus conscientes (réalisées durant les tâches *IMI* & *EXA*) présenteraient une reproduction plus précise des indices prosodiques. Si cette prédiction se retrouve dans les résultats de Sp5 dont la performance, de mieux en mieux notée au fil des tâches, semble montrer un bon contrôle prosodique ; notre prévision se trouve mise à mal par la performance de Sp1 dont certaines phrases produites en *REP* obtiennent de meilleures notes qu'en *IMI* ou en *EXA*. Ainsi, nos résultats mettent en relief une grande variabilité interlocuteur en termes de contrôle prosodique au cours des tâches.

5 Discussion & Conclusion

Cette étude visait à tester méthodologiquement une approche double d'évaluation/mesure de l'imitation des formes prosodiques. A terme, nous souhaitons développer un algorithme suffisamment robuste pour être implémenté dans un outil d'entraînement des enseignants qui leur permettrait d'évaluer la précision de leur performance prosodique.

Précédemment, le DTW, dont le terme *d* indique le coût d'alignement entre deux contours, a été utilisé par Kim (2012) en tant que mesure de la convergence phonétique. Dans notre cadre, le DTW

est principalement utilisé comme méthode d'interpolation (Rilliard et al., 2011) préliminaire au calcul de deux mesures de la similarité prosodique (initialement rapportées par Hermes, 1998). D'après ce dernier, L_2 mesure la distance perceptive entre deux contours, où une distance élevée a un facteur de poids quadratiquement plus élevé. *Zr* exprime une distance globale entre deux formes de contours, *i.e.* sa valeur indique le coût de la transformation d'un contour en un autre. Etant données leurs natures différentes, il était intéressant de les corréler tous deux aux résultats des évaluations perceptives.

Ces deux méthodes de mesure de la similarité prosodique ont obtenu une bonne corrélation avec les résultats du test AX conduits auprès de 15 auditeurs naïfs : les bonnes et les mauvaises imitations ont été repérées par l'algorithme comme par les auditeurs. La différence de corrélation entre AX et *L2/Zr* pourrait refléter la nature différente des mesures objectives. Cela étant dit, ces résultats ouvrent des perspectives intéressantes dans l'évaluation automatique de l'imitation au niveau prosodique. La question des mesures automatiques reste cependant ouverte : une étude sur plus de sujets et plus de stimuli pourrait nous renseigner sur la nécessité de conserver L_2 et/ou *Zr*. De même, de plus amples investigations seront nécessaires à la détermination d'une valeur seuil pour L_2/*Zr*, au-delà de laquelle le résultat d'un test perceptif de jugement de la similarité prosodique pourrait être estimé avec une précision suffisante.

Ces premiers résultats, encourageants, nous poussent à étendre cette approche dans de multiples directions : (1) plus d'énoncés du *CI* seront notés, objectivement et subjectivement ; (2) un corpus de phrases et de leur reproduction délexicalisée par des locuteurs sera constitué afin de tester la robustesse des mesures.

Par ailleurs, il pourrait être pertinent d'envisager l'utilisation d'autres méthodes de transformation que le DTW. Ce dernier requiert un grand nombre de valeurs de *f0*, conduisant à un temps de calcul relativement long. Parmi les méthodes de comparaison de formes passées en revue par (Veltkamp, 2001), la fonction de cumulation des angles semble pouvoir être appliquée à des contours de *f0* stylisés à partir de quelques points remarquables. Cette approche permettrait de prendre en compte explicitement certains détails des patrons prosodiques, comme les durées et les pentes et ainsi améliorer la description de ce qu'est une imitation réussie. Enfin, le raffinement de ces mesures, devrait à terme nous conduire à limiter drastiquement le recours aux tests perceptifs, en sélectionnant les mesures objectives corrélant au mieux avec des résultats perceptifs extensifs.

Les tâches accomplies pour recueillir le *CI* visaient à souligner la capacité d'imitation de locuteurs naïfs, soit, un comportement approchant celui d'un enseignant de langue étrangère sans contrôle prosodique particulier, lors de tentatives de correction phonétique. Pour certains locuteurs (par exemple, Sp1), les scores objectifs fournissent une aide au diagnostic du niveau de contrôle et de conscience prosodique, et dans une certaine mesure, du talent phonétique. Lewandowski (2012) rapporte en effet la complexité à déterminer le talent phonétique d'une personne. Ce type de mesures automatique pourrait donc être utilisé comme un indice pour détecter une composante du talent phonétique, plus précisément au niveau prosodique.

Finalement, nos perspectives de recherches se concentreront sur certains aspects spécifiques de l'entraînement à la MVT, plus particulièrement, sur la correction des indices prosodiques reproduits. A court terme, nous espérons construire une interface permettant à l'enseignant d'améliorer ses compétences d'utilisation des procédés de correction de la MVT, dans ce cas, la production de phrases délexicalisées servant à porter l'attention de l'apprenant sur la syllabification et le rythme de la langue en cours d'apprentissage.

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.

BILLIERES, M., ALAZARD, C., ASTESANO, C., & NOCAUDIE, O. (2013). Phonétique corrective en FLE : Méthode Verbo-Tonale. http://w3.uohprod.univ-tlse2.fr/UOH-PHONETIQUE-FLE/

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5:9/10*, 341‑345.

DONALD, M. (1993). *Origins of the Modern Mind - Three Stages in the Evolution of Culture & Cognition* (Reprint). Cambridge, Mass.: Harvard University Press.

DUFOUR, S., & NGUYEN, N. (2013). How much imitation is there in a shadowing task? *Frontiers in Psychology*, *4*.

GOLDINGER, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251‑279.

HERMES, D. J. (1998). Measuring the Perceptual Similarity of Pitch Contours. *Journal of Speech, Language and Hearing Research*, *41*, 73‑82.

KIM, M. (2012). *Phonetic accommodation after auditory exposure to native and nonnative speech*. NORTHWESTERN UNIVERSITY.

LEWANDOWSKI, N. (2012). *Talent in non-native phonetic convergence*. Universität Stuttgart, Stuttgart.

MICHELAS, A., & NGUYEN, N. (2011). Uncovering the Effect of Imitation on Tonal Patterns of French Accentual Phrases. In *INTERSPEECH* (p. 973–976).

MIXDORFF, H., COLE, J., & SHATTUCK-HUFNAGEL, S. (2012). Prosodic Similarity–Evidence from an Imitation Study. In *Speech Prosody 2012*.

NOCAUDIE, O., & ASTÉSANO, C. (2012). Prosodic structuring imitation in French L1 context-A first step towards correcting phonetic-prosodic features in L2 French. In *Proceedings of ISICS*. Aix-en-Provence.

PARDO, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382‑93.

PARDO, J. S. (2013). Reconciling diverse findings in studies of phonetic convergence. In *Proceedings of Meetings on Acoustics* (Vol. 19, p. 060140). Acoustical Society of America.

REVIS, J., DE LOOZE, C., & GIOVANNI, A. (2013). Vocal Flexibility and Prosodic Strategies in a Professional Impersonator. *Journal of Voice*, *27*(4), 524.e23‑524.e31.

RILLIARD, A., ALLAUZEN, A., & DE MAREÜIL, P. B. (2011). Using Dynamic Time Warping to Compute Prosodic Similarity Measures. In *INTERSPEECH* (p. 2021–2024).

VELTKAMP, R. C. (2001). Shape matching: similarity measures and algorithms (p. 188‑197). IEEE Computer Society.

ZAIONTZ, C. (2015). *Real Statistics Using Excel*. Consulté à l'adresse www.real-statistics.com

Quels tests d'intelligibilité pour évaluer les troubles de production de la parole ?

Alain Ghio[1], Laurence Giusti[2], Emilie Blanc[2], Serge Pinto[1],
Muriel Lalain[1], Danièle Robert[1, 2], Corinne Fredouille[3], Virginie Woisard[4]

(1) Université d'Aix-Marseille, 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, Toulouse, France
alain.ghio@lpl-aix.fr

RESUME

L'intelligibilité de la parole se définit comme le degré de précision avec lequel un message est compris par un auditeur. A ce titre, la perte d'intelligibilité représente souvent une plainte importante pour les patients atteints de troubles de production de la parole, puisqu'elle participe à la diminution de la qualité de vie au niveau communicationnel. Plusieurs outils existent actuellement pour évaluer l'intelligibilité mais aucun ne satisfait pleinement les contraintes cliniques. Dans une première étude, nous avons adapté au français la version 2 du Frenchay Dysarthria Assessment, un test reconnu dans le milieu anglo-saxon pour l'évaluation de locuteurs dysarthriques. Nous avons créé le corpus de mots français en nous appuyant sur les critères définis dans le FDA-2 puis nous avons testé le protocole sur une cinquantaine de locuteurs. Les résultats sont satisfaisants mais divers biais méthodologiques nous ont conduits à poursuivre notre démarche en proposant des listes de pseudo-mots apparentant le test à du décodage acoustico-phonétique.

ABSTRACT

What kind of intelligibility test to assess speech production disorders?

Speech intelligibility is defined as the precision with which a message is understood by an auditor. The loss of intelligibility is often a major complaint for patients with speech production disorders, since it contributes to reducing the quality of life. Several tools currently exist to assess intelligibility but none fully satisfies the clinical constraints. In a first study, we adapted to French language the Frenchay Dysarthria Assessment version 2, a well-known test for English speakers used for evaluating dysarthric speakers. We created the French corpus of words using criteria defined in the FDA -2 and then we tested the protocol over fifty speakers. The results are satisfactory but various methodological bias have led us to continue our efforts in proposing non sense words, which is equivalent to acoustic- phonetic decoding.

MOTS-CLES : intelligibilité, troubles de la parole, dysarthrie, perception, compréhension
KEYWORDS: intelligibility, speech disorders, dysarthria, perception, understanding

1 Introduction

1.1 Pourquoi des mesures d'intelligibilité ?

L'intelligibilité de la parole est le degré de précision avec lequel un message est compris par un individu. Si les télécommunications utilisent plutôt des mesures acoustiques qui ne font pas appel à des auditeurs et qui indiquent un indice d'intelligibilité de la parole comme, par exemple, le Speech Transmission Index (Steeneken & Houtgast, 1980), notre approche est perceptive, fondée sur des auditeurs qui écoutent des mots ou des phrases et qui répondent à une tâche de choix multiples ou de transcription. L'intelligibilité peut être évaluée dans différentes situations de communication parlée. Au niveau industriel, les tests perceptifs d'intelligibilité de la parole sont utilisés, par exemple, pour tester la qualité de communications en milieu difficile (Ghio et al., 2006). Dans des cas de malentendance, l'intelligibilité peut être évaluée pour vérifier les capacités perceptives d'un auditeur avec la méthode de l'audiométrie vocale (Lafon et al., 1965). Elle peut enfin être estimée chez des locuteurs atteints de troubles de production de la parole (TPP), une perte d'intelligibilité représentant souvent une plainte importante de la part de ces patients puisqu'elle participe à la diminution de la qualité de vie au niveau communicationnel. C'est dans ce dernier cadre que nous situons notre travail.

Dans un contexte orthophonique ou phoniatrique, la mesure de l'intelligibilité est généralement utilisée pour évaluer le degré de sévérité du trouble de la parole, une équivalence étant admise entre une baisse d'intelligibilité et une augmentation de la sévérité du dysfonctionnement. Cela peut concerner à la fois les troubles touchant directement les organes de la phono-articulation (post-chirurgie du cancer, fentes palatines…) mais aussi les troubles d'origine neurologique (dysarthries). Cette mesure d'intelligibilité est alors un élément du bilan orthophonique ou phoniatrique permettant d'orienter la prise en charge du patient (travail de compensation des troubles articulatoires, proposition de communication alternative) et d'apprécier un résultat thérapeutique (chirurgie, traitement pharmaceutique, rééducation). Dans ce contexte, l'intelligibilité se mesure habituellement de la façon suivante : le patient lit une liste de mots et l'examinateur transcrit ce qu'il comprend. Les transcriptions sont ensuite confrontées à la liste lue et le score d'intelligibilité est le pourcentage d'éléments correctement reconnus. Si l'examinateur est généralement unique en consultation clinique, le nombre d'auditeurs peut être étendu à un groupe plus important (jury d'écoute) dans des contextes de recherche comme notre étude.

1.2 L'intelligibilité et la compréhensibilité de la parole

La perception de la parole est un mécanisme qui intègre à la fois un flux ascendant d'informations provenant du signal (décodage acoustico-phonétique) et un flux descendant à partir des informations de haut niveau détenues par l'auditeur (lexicalité, contexte de la situation de communication, environnement et connaissances des communicants). Lorsque nous entendons 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 (Ganong, 1980) et nous pouvons prédire qu'en français une séquence prononcée [tisk] sera perçue /disk/ en référence au mot « disque », et inversement, une séquence [daʃ] sera perçue /taʃ/ en référence au mot « tache ». De ce fait, quand un auditeur communique avec un locuteur ayant des troubles de la production de la parole, il sollicite particulièrement ces stratégies de compensations qui intègrent ces aspects de lexicalité mais aussi les effets de 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 et même la familiarité avec le type de trouble de la parole.

Dans un cadre d'évaluation des troubles de la production de la parole, ces mécanismes top-down peuvent s'avérer gênants pour mesurer le degré de 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'éloigne de l'évaluation de l'altération chez le locuteur en se plaçant sur le versant de l'invalidité, voire de son potentiel handicap au sens de la terminologie de l'OMS (Rossignol, 2007). 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).

2 L'adaptation en français du FDA-2

2.1 Les tests d'intelligibilité en clinique

Pour le français, le test de diagnostic par paires minimales de Peckels & Rossi (1973) est l'un des plus anciens et des plus aboutis mais il est peu utilisé en contexte clinique car il nécessite la production de 216 mots, ce qui est trop contraignant pour un patient atteint de troubles de la parole. Le *Single Word Intelligibility Test* (Kent, 1989) a été adapté et traduit en français par Gentil (1992) puis repris par Auzou en 1998 dans l'ECD et par Crochemore & Vannier en 2001 pour devenir le Test Phonétique d'Intelligibilité de la Batterie d'Evaluation Clinique de la Dysarthrie (Auzou et Rolland-Monnoury, 2006). Dans cet outil standardisé fréquemment utilisé en France pour évaluer la dysarthrie dans son ensemble, on retrouve un autre test d'intelligibilité plus court (production de 10 mots isolés), qui est directement adapté d'un test anglo-saxon, le *Frenchay Dysarthria Assessment* (Enderby, 1983). Une nouvelle version de ce test anglo-saxon a été proposée en 2008 (FDA-2, Enderby et Palmer, 2008). Nous nous sommes alors donnés comme objectif d'adapter ce test d'intelligibilité à la langue française (Blanc & Giusti, 2014).

2.2 Méthode et corpus

La première étape a consisté en la création d'un nouveau corpus de 116 mots (Source : lexique-3, www.lexique.org) en s'appuyant sur des critères définis dans le FDA-2. Nous avons notamment contrôlé :

- la fréquence des mots (supérieure à 10 apparitions par million de mots),
- la place des consonnes en position initiale, médiane et finale (nos phonèmes de référence sont toutes les consonnes de la langue française et les groupes consonantiques les plus fréquents en position initiale)
- des listes de mots avec 1, 2, 3 ou 4 syllabes

A ces critères issus du FDA-2, nous avons ajouté d'autres critères pour la sélection des mots de notre corpus :

- une structure phonotactique constante à l'intérieur de chacune des listes des mots courts.
- une variété des sons vocaliques au sein des mots d'une liste. Ainsi, nous limitons le risque que l'examinateur identifie le mot à la seule discrimination d'une voyelle (ex : dans la version 2006 de la BECD, « mouche » est le seul mot de la liste qui contient le phonème [u])
- un contrôle du phonème final pour les mots bisyllabiques
- un point d'unicité repoussé au maximum pour les mots longs. Le point d'unicité est le point à partir duquel un mot se distingue d'un autre. Ainsi, pour le mot « vérité », par exemple, ce n'est qu'à partir de la troisième syllabe que l'on peut être sûr qu'il ne s'agit pas du mot « véridique », « véritable »...

Par la définition de ces critères supplémentaires, notre objectif est double : affiner l'équilibre phonétique et réduire l'effet d'apprentissage de l'examinateur. En effet, si plusieurs mots utilisent les mêmes sons vocaliques, l'auditeur restaurera moins facilement un mot mal prononcé, même s'il connaît le corpus de mots. L'intelligibilité du locuteur sera ainsi davantage mise à l'épreuve que la capacité de restauration de l'auditeur. La liste complète est présentée en Table 1.

	Lieu	*Monosyll.*	*Bisyll.*	*Mono+ schwa*	*Trisyllabique*	*Quadrisyllabique*
	d'art.	*Initiale*	*Médiane*	*Finale*	*Initiale*	*Initiale*
[p]	bilabiale	pain	appel	soupe	politique	population
[b]		bar	habits	robe	bâtiment	bénédiction
[m]		mer	amour	pomme	magasin	majorité
[f]	fricative	feu	enfin	gaffe	fatigue	fidélité
[v]		vin	envie	rêve	vérité	/
[t]	dentale	temps	autour	vite	téléphone	télévision
[d]		donc	aider	code	décider	débarrasser
[n]		non	année	bonne	numéro	normalement
[s]	alvéolaire	sac	aussi	douce	solitude	sécurité
[z]		zut	hasard	chaise	/	/
[ʃ]	palatale	cher	échelle*	bouche	charité	/
[ʒ]		jour	agir	rouge	général	génération
[k]	vélaire	cœur	écart	chèque	qualité	conversation
[g]		goût	égal	bague	gouverneur	gouvernement
[l]	Liquide	lac	aller	balle	légitime	laboratoire
[r]		rue	héros	père	relation	récupérer
[w]	semi-voyelle	oui	avoir	/	/	/
[j]		hier	ancien	fille	/	/
[ui]		huit	enfuir	/	/	/
[gn]		/	agneau	ligne	/	/

Monosyllabiques avec groupe consonantique à l'initiale :

[pl] plat	[pr] prêt	[bl] bleu	[br] bref	[fl] fleur	[fr] front	[tr] train	[dr] draps
[kl] clair	[kr] cri	[gl] glace*	[gr] grand	[sp] sport	[st] stop	[sk] ski	[vr] vrai

TABLE 1 : Corpus de mots français, adapté du FDA-2

2.3 Validation

Une fois la liste constituée, un test de validation a été entrepris auprès de locuteurs sains, grâce à l'enregistrement de cinquante personnes qui ont lu chacune 10 mots au hasard dans la liste car tel est le protocole standard. Ces stimuli ont été ensuite présentés à dix-huit auditeurs dont la tâche était de transcrire le mot entendu. Afin de démontrer que le test était sensible à la dégradation du signal de parole, chaque item a été présenté avec et sans bruit (bruit blanc avec niveau de bruit à 60% du niveau de signal). Les résultats montrent que le test est valide pour les locuteurs sains car le score d'intelligibilité est de 96% en condition normale (sans bruit). La sensibilité à la dégradation est aussi observée dans la mesure où les scores sont significativement abaissés pour les stimuli dégradés par du bruit (score de 54 % en moyenne). Nous observons un effet significatif de la longueur des mots sur le taux d'intelligibilité en condition de parole dégradée. En effet, le score passe de 39% pour des mots monosyllabiques à 84% pour les mots quadrisyllabiques (figure 1, gauche). Ces résultats sont expliqués par un effet de quantité d'information maximisant ou pas la restauration phonémique comme cela a été montré dans les travaux de Samuel (1981). Nous suggérons désormais de procéder à un tirage au sort par catégorie : 7 mots tirés au sort parmi les mots courts, 3 mots tirés au sort parmi les mots longs. De plus, nous observons une augmentation significative du score d'intelligibilité qui passe de 45% en début de test pour atteindre 60% en fin de séance pour la parole dégradée (figure 1, droite). Nous expliquons ce résultat par le fait que les auditeurs développent un degré d'expertise d'écoute au cours de l'exercice et parviennent de mieux en mieux à restaurer un message dégradé. De plus, il existe un effet d'apprentissage au cours du temps : les mêmes mots (prononcés par différents locuteurs) revenant plusieurs fois, l'auditeur fini par en reconnaître

certains comme faisant partie du corpus, les identifiant de mieux en mieux, même dans le bruit. Ce taux d'intelligibilité qui s'améliore dans le temps illustre bien la problématique clinique liée à l'habituation du thérapeute au test, ce qui nous a conduits à une nouvelle alternative présentée ci-après.

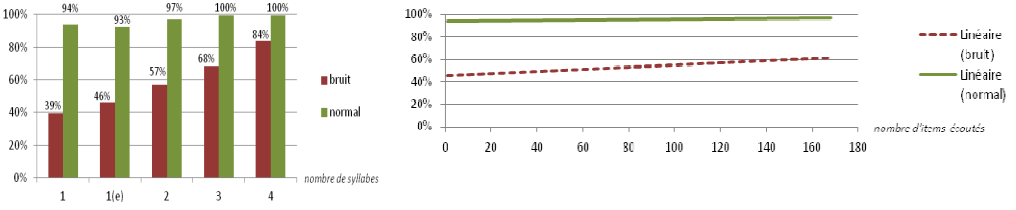


Figure 1 : scores d'intelligibilité en fonction de la longueur des mots (à gauche) ou du nombre d'items écoutés (à droite) pour de la parole normale ou bruitée

3 Intelligibilité sous forme de décodage acoustico-phonétique

3.1 Pourquoi du décodage acoustico-phonétique ?

Les limitations des tests d'intelligibilité effectués sur des locuteurs ayant des troubles de production de la parole 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. Le biais lié à cette connaissance et donc à la forte influence des mécanismes perceptifs descendant est un score d'intelligibilité surévalué car la restauration phonémique de l'auditeur rend opaque les distorsions de production. Imaginons un locuteur avec des difficultés vocales liées à une opération au niveau laryngé qui prononce la séquence « tocteur » de façon dévoisée, l'auditeur reconnaitra l'item « docteur », ce qui comptera comme une bonne réponse et donc comme une absence d'altération, ce qui n'est pas le cas. Un tel problème d'intelligibilité surévaluée ne rendra pas compte du réel déficit communicationnel d'une part et d'autre part entraînera un biais possible dans l'élaboration du projet thérapeutique.

Il existe différentes méthodes pour neutraliser ces effets. La première solution est d'annuler l'aspect prédictif des items, ce qui est le cas dans les tests de paires minimales (Peckels & Rossi, 1973). Dans ce type de tests, les mots « coule », « poule », « boule », « moule » sont équiprobables et permettent de tester le fonctionnement laryngé (« poule » vs « boule »), le fonctionnement vélaire (« moule » vs « boule »), le fonctionnement articulatoire (« coule » vs « poule »)… L'inconvénient de ce test est qu'il ne teste que les consonnes à l'initiale, ce qui restreint le pouvoir d'évaluation. Une autre façon de diminuer l'effet prédictif est l'usage d'un grand lexique de plusieurs milliers de mots mais l'auditeur reste dépendant des effets de fréquences lexicales. De plus, il est difficile dans un tel contexte d'obtenir des listes équivalentes, c'est-à-dire contenant globalement les mêmes quantités d'unités phonétiques et des structures similaires.

La solution que nous avons adoptée 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 qui s'apparente à du décodage acoustico-phonétique suivi d'une transcription écrite.

3.2 Matériel

Pour la construction du matériel linguistique, nous avons utilisé les données fournies par www.lexique.org. Les pseudo-mots ont été construits de la façon suivante au niveau phonétique:

- Ils ont la structure phonotactique du type $C(C)_1V_1$ $C(C)_2V_2$ où $C(C)_i$ est une consonne isolée ou un groupe consonantique
- $C_1 \in$ {p t k b d g v z ʒ f s ʃ r l m n ñ yod} 18 éléments
- $CC_1 \in$ { pr tr kr gr br fr pl kl fl st bl sk sp gl dr ps }
  qui sont les 16 groupes consonantiques en position initiale les plus fréquents en français d'après les données de « lexique.org ». Ces 16 éléments représentent 97% des groupes consonantiques en position initiale.
- $V_1 \in$ { a i y u o e ã ~ɛ } 8 éléments
  Nous considérons l'unité « o » comme l'archiphonème des variantes ouvertes et fermées.
  Nous considérons l'unité « e » comme l'archiphonème des variantes [e], [ɛ], [oe], [Ø].
- $C_2 \in$ {p t k b d g v z ʒ f s ʃ r l m n ñ yod} 18 éléments
- $CC_2 \in$ { st ks rd rs kt rn pl gr dr kl rj lt rv vr gz rp tr rt bl rm pr kr sk br sp rk fr fl rb gl ps pt}
  qui sont les 32 groupes consonantiques en position intervocalique les plus fréquents en français d'après les données de « lexique.org ». Ces 32 éléments représentent 87% des groupes consonantiques en position intervocalique.
- $V_2 \in$ { a i y u o ã ~ɛ } 7 éléments
  En position finale, nous avons éliminé l'unité « e » qui une fois écrite, pourrait être considérée comme un « schwa » et donc poser des problèmes de variabilité interindividuelle, notamment régionale.

Une telle combinatoire permet de générer (18 C_1 + 16 CC_1) * 8 V_1 * (18 C_2 + 32 CC_2) * 8 V_2 = 95200 formes. Sur ces 95200 formes, nous avons éliminé 5854 formes qui étaient des mots du lexique, ce qui nous laisse une liste de 89346 pseudos-mots. Nous remarquons au passage que seulement 6% des possibilités d'une telle structure ont émergé au niveau lexical pour le français, ce qui met en évidence le faible taux de remplissage de l'espace des structures phonético-lexicales.

3.3 La conversion phonème-graphème

Notre liste de pseudo-mots ayant été construite sur des critères phonologiques, nous nous sommes heurtés à une problématique non triviale de conversion phonème-graphème de façon à fournir des formes orthographiques déterministes, compatibles avec les règles générales du français, faciles à lire et sans ambigüité. Nous avons ainsi un certain nombre de règles comme ci-dessous :

[ã+b|p]	⇨ « amb|p »	sinon	[ã] ⇨ « an »
[v_{oy} + s + v_{oy}]	⇨ « v_{oy} ss v_{oy} »		
[k+i|e]	⇨ « qui|e »	sinon	[k] ⇨ « c »
[g+i|e]	⇨ « gui|e »		
[a|u|e+yod]	⇨ « a|ou|eille » , [i+yod]	⇨ « iy »	

3.4 Méthode

Si dans le FDA-2, le protocole prévoyait uniquement la lecture de 10 mots par locuteur, nous proposons d'augmenter cette quantité à 52 items, les deux premiers étant des essais d'entrainement non comptabilisés dans les résultats. Ce nombre de 52 est intéressant car il permet de constituer des listes dans lesquelles systématiquement :

- Apparaissent deux fois chaque consonne C_1 à l'initiale et une fois l'un des 16 groupes consonantiques CC_1 (2 * 18 + 16 = 52)
- Chaque voyelle V1 apparait au moins 6 fois et (a ; i ; ou, e) une fois de plus (8*6 + 4 = 52)
- Apparaissent deux fois chaque consonne C_2 en inter-vocalique et la moitié des 32 groupes consonantiques CC_2 (2 * 18 + 16 = 52)
- Chaque voyelle V2 apparait au moins 7 fois et (a ; i ; ou) une fois de plus (7*7 + 3 = 52)

Un algorithme intégrant ces contraintes et piochant dans la liste des 89346 pseudos-mots permet d'obtenir des listes de logatomes différents dont la structure phonotactique et le contenu phonémique demeurent néanmoins identiques.

crampant	fevo	quinfant	flaspou	plouniant
troucha	suptu	vabla	baillu	ratri
rougli	nougu	touflant	griti	ninrtin
dibro	yango	zucrou	quebo	gavi
pufru	scuchu	psoussa	chanjin	bijo
blouillu	fampsi	madin	lupou	tanli
niascu	pimprant	climbou	storquin	brori
chansin	jindou	spucou	glima	prinrmo
gomou	droto	yezant	vefin	zelin
jezant	dorba	nioniou	mera	frina
lina	siqui			

TABLE 2 : Un exemple de liste de pseudo-mots

3.5 Application

Ce protocole est en cours d'utilisation 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.

Afin de faciliter l'élicitation du corpus, chaque pseudo-mot est à la fois affiché à l'écran et produit oralement par un expérimentateur. Plus récemment, nous avons commencé à tester la possibilité d'utiliser de la synthèse vocale pour remplacer le côté variable de l'expérimentateur. Même si la tâche est difficile, elle s'avère parfaitement faisable et s'avère bien plus sensible à la dégradation que les tâches classiques d'intelligibilité. Les résultats de ce travail seront présentés ultérieurement.

Conclusion

Les tests classiques d'intelligibilité de la parole ne sont pas forcément adaptés à l'évaluation des troubles de la production de la parole. En effet, l'auditeur mettant en place des stratégies de compensation, le résultat ne reflète que partiellement l'état de dysfonctionnement du locuteur. Pour s'en rapprocher, le décodage acoustico-phonétique de pseudo-mots pourrait s'avérer préférable, surtout si d'autres tests comme de la compréhension globale sont appliqués en complément.

Remerciements

Une partie de 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.

Références

AUZOU P, ROLLAND-MONNOURY V. (2006), Batterie d'évaluation de la dysarthrie, *1st ed. Isbergues: Ortho Edition.*

AUZOU P, OZSANCAK C, JAN M, LEONARDON S, MENARD JF, GAILLARD MJ, EUSTACHE F, HANNEQUIN D. Evaluation clinique de la dysarthrie : présentation et validation d'une méthode. *Rev Neurol (Paris)*. 1998;154(6-7):523-530.

BLANC E, GIUSTI L. (2014) Evaluation de l'intelligibilité de la parole dans les dysarthries : adaptation en français de la version révisée du FDA2, Certificat d'orthophonie, Marseille.

CROCHEMORE E, VANNIER F. (2001) Analyse phonétique de la parole dysarthrique. In : *Les Dysarthries*. Auzou P., Özsancak C., Brun V. (Eds) Masson, Paris,: 71-82.

ENDERBY P. (1983) Frenchay Dysarthria Assessment. *1st ed. San Diego: College-Hill Press.*

ENDERBY P, PALMER R. (2008*)* FDA-2: Frenchay Dysarthria Assessment. *2nd ed.Tex.: Pro-Ed.*

GANONG W. (1980). Phonetic categorization in auditory word perception. Journal of Experimental Psychology. *Human perception and performance*, 6 (1), 110-125.

GENTIL M. Phonetic intelligibility testing in dysarthria for the use of French language clinicians. *Clin Linguist Phon*. 1992;6(3):179-189.

GHIO A., MEYNADIER Y., TESTON B., LOCCO J., CLAIRET S., ESPESSER R., MEUNIER C., VINCENT-MARLIEN I., DENIAUD C. (2006). Peut-on parler sous l'eau avec un embout de détendeur ? Etude articulatoire et perceptive. *Actes, Journées d'Etude sur la Parole (JEP),* p. 379-382

KENT RD, WEISMER G, KENT JF, ROSENBEK JC. (1989) Toward phonetic intelligibility testing in dysarthria. *J Speech Hear Disord*. 54(4):482-499.

LAFON JC, MORGAN A, GAUTHIER J (1965) L'intervalle de confiance des mesures audiométriques vocales, *Int. Audiol*., 4, 94-96

PECKELS J, ROSSI M. (1973) Le Test de diagnostic par paires minimales : adaptation au français du Diagnostic Rhyme Test de Voiers. *Revue d'acoustique*; 27:245-262.

ROSSIGNOL C. (2007). Classifications internationales des altérations corporelles, dysfonctionnements et handicaps. Pour une clarification des concepts. *Actes des Entretiens de Bichat, Orthophonie*. Paris: Expansion Formation Editions, p. 62-69.

SAMUEL A. Phonemic restoration: insights from a new methodology. *J Exp Psychol Gen*. 1981;110:474-494.

STEENEKEN, H. J. M., & HOUTGAST, T. (1980) A physical method for measuring speech-transmission quality, *Journal of the Acoustical Society of America*, 67, 318–326

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.

Réalisation phonétique et contraste phonologique marginal : une étude automatique des voyelles du roumain

Ioana Vasilescu[1] Margaret E. L. Renwick[2] Camille Dutrey[3,4]
Lori Lamel[1] Bianca Vieru[5]
(1) LIMSI, CNRS, Université Paris-Saclay , Bât. 508, Orsay, France
(2) University of Georgia, Athens GA 30602, USA
(3) LPP, CNRS, Université Paris 3, 19 rue des Bernardins, 75005 Paris, France
(4) Laboratoire National de Métrologie et d'Essais, 29 avenue Roger Hennequin, 78190 Trappes, France
(5) Vocapia, 28 Rue Jean Rostand, 91400 Orsay, France
ioana,lamel@limsi.fr, mrenwick@uga.edu, camille.dutrey@lne.fr,
vieru@vocapia.fr

RÉSUMÉ

Cet article est dédié à l'analyse acoustique des voyelles du roumain : des productions en parole continue sont comparées à des prononciations "de laboratoire". Les objectifs sont : (1) décrire les traits acoustiques des voyelles en fonction du style de parole ; (2) estimer la relation entre traits acoustiques et contrastes phonémiques de la langue ; (3) estimer dans quelle mesure l'étude de l'oral apporte des éclairages au sujet des attributs phonémiques des voyelles centrales [ʌ] et [ɨ], dont le statut (phonèmes vs allophones) est controversé. Nous montrons que les traits acoustiques sont comparables pour la parole journalistique vs contrôlée pour l'ensemble de l'inventaire *sauf* [ʌ] et [ɨ]. Dans la parole contrôlée [ʌ] et [ɨ] sont distinctes, mais confondues en faveur du timbre [ʌ] à l'oral. La confusion de timbres n'est pas source d'inintelligibilité car [ʌ] et [ɨ] sont en distribution quasi-complémentaire. Ce résultat apporte des éclairages sur la question du contraste phonémique graduel et marginal (Goldsmith, 1995; Scobbie & Stuart-Smith, 2008; Hall, 2013).

ABSTRACT

Phonetic realization and marginal phonemic contrast : an automatic study of the Romanian vowels

This paper compares acoustic properties of Romanian vowels in continuous vs laboratory speech. We aim to : (1) describe acoustic features of Romanian vowels as function of the speaking style ; (2) estimate the impact of speaking style on language-specific phonemic contrasts ; (3) correlate acoustic properties of central vowels [ʌ] and [ɨ] in continous speech with their phonemic status. Results show that while phonologically the central vowels have different representations, this contrast is severely diminshed in continous speech, leading to a phonetic near-merger. The phonological competence is not realized in continuous speech, where performance permits considerable overlap among central vowels. The vowels' nearly complementary distribution circumvents intelligibility loss. The results provide support to the hypothesis of gradient (instead of categorical) and marginal phonemic contrast (Goldsmith, 1995; Scobbie & Stuart-Smith, 2008; Hall, 2013).

MOTS-CLÉS : Système vocalique roumain, voyelles centrales, parole continue, contraste phonologique marginal.

KEYWORDS: Romanian vowels, central vowels, continuous speech, marginal phonemic contrast.

1 Introduction

La parole continue est un phénomène hautement variable dans le temps et dans l'espace. Des facteurs tels que les locuteurs et leurs caractéristiques physiologiques, les contextes communicationnels ou encore les conditions d'enregistrement concourent à sa variabilité (Benzeghiba *et al.*, 2007; Adda-Decker, 2006). Cette dernière décennie est née une méthodologie conjuguant traitement automatique et analyse acoustico-prosodique des marques de la variation linguistique dans de grands corpora oraux. Les travaux menés dans ce cadre ont un double objectif : améliorer les systèmes automatiques et répondre à des questions linguistiques en s'appuyant sur des données "tout venant". Pour les linguistes, il s'agit d'une approche novatrice permettant de confronter à l'existant des hypothèses construites "en laboratoire". Le travail présenté ici s'inscrit dans cette lignée se penchant sur le roumain, une langue pour laquelle les avancées en traitement automatique de la parole sont récentes et dont la forme orale spontanée est peu décrite (Trandabat *et al.*, 2012).

Le roumain est une langue romane, issue de la branche orientale de cet idiome. Il est langue maternelle pour environ 29 millions de locuteurs à travers le monde et langue officielle de deux pays : la Roumanie et la République Moldave. Des raisons géo-politiques expliquent son évolution isolée par rapport aux autres membres de la famille latine. L'influence slave est souvent évoquée en tant que spécificité de la langue. La phonétique et phonologie du roumain bénéficient d'une longue tradition scientifique, les linguistes roumains ayant souvent associé leur analyses à des questions plus larges, historico-politiques (Rosetti, 1986). Parmi les sujets débattus nous pouvons noter l'origine et la place dans le système des voyelles centrales [ʌ] [1] et [ɨ], qui représentent une innovation par rapport au latin (Avram, 2012) ou bien l'existence d'une palatalisation et labialisation phonémiques, marques de l'influence slave (Avram, 1992; Petrovici, 1956). Cependant toutes ces démarches restent théoriques ne s'appuyant sur l'oral que dans la mesure où il s'agit d'une perspective dialectologique. Plus récemment, des analyses phonétiques (acoustiques et perceptives) sont menées à partir de données acquises dans des conditions "de laboratoire" (Spinu *et al.*, 2012; Chitoran, 2002a; Renwick, 2014).

Dans cette étude, nous nous intéressons au système vocalique roumain, à la fréquence et aux caractéristiques acoustiques des voyelles dans la langue parlée, ainsi qu'aux implications systémiques des contrastes observés. L'inventaire vocalique roumain comporte sept monophtongues : [i], [e], [u], [o], [a], [ʌ], [ɨ] (Graur & Rosetti, 1938; Chitoran, 2002b). Nous comparons leur réalisation dans la parole continue, illustrée par un corpus journalistique, à des prononciations canoniques, enregistrées en laboratoire (Renwick, 2014). Nous étudions tout particulièrement le rapport entre réalisation effective et statut phonémique des voyelles centrales [ʌ] et [ɨ]. [ʌ] et [ɨ] sont historiquement des allophones, le statut de phonèmes étant acquis tardivement grâce à l'émergence de quelques paires minimales (par ex., *rău* [rʌw] "méchant" vs *râu* [rɨw] "rivière"). La distribution est quasi-complémentaire, [ʌ] participe aux flexions nominale et verbale (par ex., marque de l'article indéfini *fata/fată* [fata]/[fatʌ] "la fille/une fille" et du subjonctif verbal *să facă* [sʌ.fakʌ] "il fasse"), tandis que [ɨ] apparaît souvent en contexte nasal dans de nombreux mots outils (par ex., *în* [ɨn] prep. "dans") (Avram, 2012). Néanmoins, la réalisation en tant que timbres distincts dans la parole de laboratoire soutient leur statut de phonèmes "pleins" (Renwick, 2014). Nous discutons ici leur fréquence, la distribution dans le corpus et les traits acoustiques dans la parole continue et faisons l'hypothèse que les deux phonèmes illustrent la notion de contraste phonologique marginal (Goldsmith, 1995; Scobbie & Stuart-Smith, 2008; Hall, 2013). Selon (Goldsmith, 1995) le contraste phonémique, traditionnellement décrit comme

1. Deux transcriptions sont proposées pour la voyelle centrale d'aperture moyenne : [ə] et [ʌ]. Nous adoptons la dernière afin d'éviter les confusions potentielles avec une voyelle de type *schwa*, dont la voyelle roumaine ne partage pas les traits phonologiques.

catégoriel, suit un continuum allant d'une opposition proprement-dite à des différents degrés de neutralisation contextuelle. Ces relations de type marginal correspondent à des sous-catégories du contraste phonémique, parmi lesquelles la classe des "sons à peine contrastifs" qui pourrait concerner le contraste [ʌ]/[ɨ] en roumain.

La section suivante (section 2) décrit les données utilisées et la méthodologie d'analyse. La section 3 décrit les résultats : nous présentons les caractéristiques de fréquence et les spécificités acoustiques des voyelles en fonction du style de parole. Les voyelles centrales d'aperture moyenne et fermée, dont la distribution et les caractéristiques acoustiques remettent en question leur statut de phonèmes "pleins" sont tout particulièrement considérées. Enfin, la dernière section est dédiée aux conclusions (section 4).

2 Corpora et methodologie

2.1 Corpora

Nous utilisons deux types de données que nous associons à trois styles de parole : journalistique semi-préparé (informations lues ou discours politique préparé) (*Corpus_prep*), journalistique spontané (débats télévisés) (*Corpus_spont*) et mots prononcés dans des phrases porteuses, enregistrés en laboratoire (*Corpus_lab*).
Le corpus journalistique à été acquis lors de la réalisation d'un premier système de reconnaissance automatique de la parole continue pour le roumain (Vasilescu *et al.*, 2014). Nous utilisons ici les données de développement et d'évaluation qui totalisent environ 7h et qui bénéficient de transcriptions manuelles exactes. Ces données sont issues d'émissions radio et télé-diffusées variées (Euranet, RFI Journal, RRA - Radio România Actualităţi, Antena3). Les enregistrements contiennent des locuteurs masculins et féminins, la proportion des premiers étant plus importante (environ 70%). Le nombre de locuteurs varie selon les sources, allant de 3 (Euranet) à 24 (Antena 3). Les sources couvrent globalement la variété standard de la langue, fondée sur le dialecte daco-roumain parlé dans le Sud de la Roumanie, qui constitue la base de la langue littéraire. Le style de parole est caractéristique du discours public, lu et/ou préparé, une exception étant constituée par la source Antena 3 où il s'agit d'un débat faisant intervenir des interlocuteurs de différents horizons socio-professionnels. De ce fait nous établissons d'emblée une dichotomie entre les sources de type "parole (semi-)préparée" (Euranet, RFI Journal, RRA - Radio Romania Actualităţi) (*Corpus_prep*) et "parole spontanée" (Antena 3) (*Corpus_spont*.

	Corpus_prep	*Corpus_spont*
Durée	3h32'	3h37'
Nb sources	3	1
Nb mots	31 299	24 997
Nb mots distincts	6 400	4 563
Nb locuteurs	93	48

TABLEAU 1 – Description générale des données journalistiques en fonction du style de parole : parole préparée (*Corpus_prep*) *vs* parole spontanée (*Corpus_spont*).

Corpus_lab inclut les sept voyelles du roumain en syllabes accentuées et non accentuées, extraites de mots prononcés par 18 locuteurs, 3 hommes et 15 femmes. Chaque mot est lu trois fois dans une phrase porteuse. Les prononciations sont standard et canoniques : des analyses acoustiques et une expérimentation perceptive soutiennent des réalisations distinctes des sept monophtongues (Renwick, 2014).

2.2 Méthodologie

Le corpus journalistique (*Corpus_prep* et *Corpus_spont*) a été traité automatiquement. Un alignement forcé des transcriptions manuelles de référence a été réalisé en utilisant les modéles acoustiques et le dictionnaire utilisés pour le système de transcription automatique de la parole décrit dans (Vasilescu *et al.*, 2014), générant une segmentation en mots et phonèmes. Par la suite des paramètres acoustiques (fréquence fondamentale, durée, formants F1, F2) ont été extraits automatiquement avec Praat (Boersma & Weenink, 2001) selon la procédure décrite dans (Gendrot & Adda-Decker, 2005). Le corpus contient 296.042 segments, dont 125.501 voyelles. Le taux de voisement (calculé comme rapport entre le nombre de trames voisées et le nombre total de trames du segment) a été estimé pour chaque item. Seulement les voyelles ayant un taux de voisement supérieur à 40% ont été retenues (120.479 voyelles). Un filtrage acoustique strict destiné à éliminer les données aberrantes (résultat de conditions d'enregistrement défectueuses, de prononciations approximatives et/ou d'erreurs d'alignement) est également appliqué et des items sont rejetés sur la base de la distance de Mahalanobis (Mahalanobis, 1971). Nous obtenons *in fine* 104.456 items vocaliques qui représentent 83.2% des données vocaliques, soit 60.646 pour *Corpus_prep* et 43.810 pour *Corpus_spont*. Les valeurs formantiques sont normalisées par locuteur (Lobanov, 1971).

Pour ce qui est de (*Corpus_lab*), les mots ont été segmentés et vérifiés manuellement et les valeurs centrales des deux premiers formants F1, F2 ont été extraites automatiquement avec Praat. Les données ont généré 5.261 valeurs utilisées dans l'analyse [2].

3 Résultats

3.1 Fréquence des voyelles dans les données journalistiques

Nous avons estimé la fréquence des voyelles dans les données journalistiques (*Corpus_prep* et *Corpus_spont*). Cette analyse a comme premier objectif d'offrir une vue d'ensemble de l'usage effectif des voyelles dans la langue parlée. Un second objectif est d'observer si des éléments de fréquence peuvent être associés aux conditionnement phonologique des voyelles centrales [ʌ] et [ɨ]. Le corpus journalistique totalise 56k items lexicaux dont 9.032 mots distincts.

La figure 1 montre la fréquence dans le corpus des sept monophtongues. Les voyelles [a], [e] et [i] représentent les éléments les plus fréquents confirmant des estimations antérieures faites à partir de données écrites (liste de mots) (Renwick, 2014). Les voyelles centrales [ʌ] et [ɨ] totalisent moins de 10% des voyelles du corpus. Ce résultat montre que les deux voyelles sont les éléments les plus

2. Le roumain est une langue à accent lexical. Si l'accent n'affecte pas les paramètres formantiques des voyelles enregistrées en laboratoire, il se traduit néanmoins par des durées plus longues des voyelles accentuées (Renwick, 2014). Dans cette étude préliminaire les voyelles sont considérées indépendamment de leur position par rapport à l'accent, souvent associé à la pénultième syllabe du mot. L'influence de l'accent reste néanmoins un aspect à prendre en compte dans des études futures.

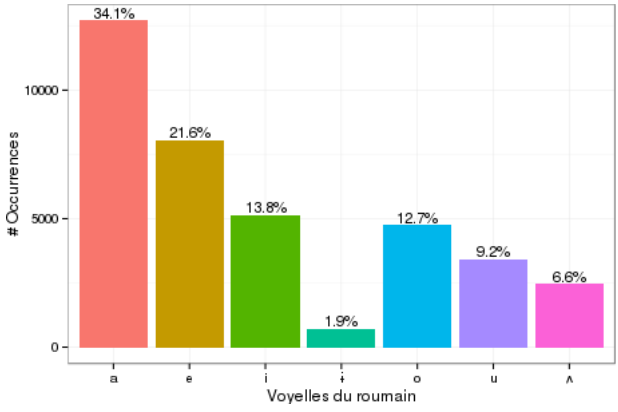


FIGURE 1 – Fréquence des voyelles dans les données journalistiques (*Corpus_prep* et *Corpus_spont*).

rares du système malgré leur usage dans de nombreux morphèmes grammaticaux et mots outils. Ce pourcentage est légèrement inférieur à celui mesuré sur *Corpus_lab* (Renwick, 2014).

Afin d'avoir un éclairage sur les mots contenant les voyelles [ʌ] et [ɨ], nous avons estimé leur contextes d'occurrence les plus fréquents. Le tableau 2 met en évidence une distribution fortement complémentaire : [ɨ] apparaît surtout dans le mot "România", fréquence expliquée par le style de parole (journalistique) et en position initiale de mot, où [ʌ] n'apparaît que très rarement. Inversement, [ʌ] se trouve souvent en position finale de mot, notamment en tant que marque du subjonctif verbal (*să facă* [sʌ.fakʌ] "il fasse"), où [ɨ] n'apparaît jamais. La forme nominale indéfinie, a priori un cas de figure pouvant contribuer à la productivité de [ʌ], n'est pas fréquente dans nos données. Les deux voyelles peuvent précéder une consonne : [ɨ] se retrouve majoritairement devant une consonne nasale, conséquence de son émergence en tant qu'effet de la pré-nasalisation, ou devant la liquide [r] en position initiale ou intra-mot ce qui correspond à la quasi-totalité des contexte d'occurrence de la voyelle, tandis que [ʌ] peut être suivie par d'autres consonnes. Par opposition [ʌ] apparaît le plus souvent en position post-consonantique, en tant que marque du subjonctif verbal ou des formes nominales indéfinies.

Contexte	/ɨ/		/ʌ/	
C**V**C	30.5%	România	26.6%	astăzi
C**V**V	1.3%	mâine	1.8%	său
V**V**C	0.3%	neîncadrabili	2.8%	două mii doisprezece
#**V**C	67.2%	în	0.4%	ăsta
#**V**V	0.7%	îi	0.1%	ăia
C**V**#	0.0%	–	65.8%	să
V**V**#	0.0%	–	2.5%	două

TABLEAU 2 – Distribution des voyelles centrales /ɨ/ et /ʌ/ selon la position dans le mot : intra-mot (C**V**C, V**V**C, C**V**V), initiale (#**V**C, #**V**V) et finale (C**V**#, V**V**#). Exemple du mot le plus fréquent pour chacun des cas.

3.2 Caractéristiques acoustiques

Nous nous sommes intéressées aux paramètres **timbre** et **durée** des voyelles que nous avons analysées en fonction du style de parole. La Figure 2 montre la distribution dans l'espace des formants F1/F2 des sept monophtongues de *Corpus_lab*, *Corpus_prep* et *Corpus_spont*. Les dispersions concernent les locuteurs féminins, mieux représentés dans *Corpus_lab*, néanmoins des effets similaires ont été observés chez les locuteurs masculins. *Corpus_lab* présente les espaces vocaliques les plus distincts, par opposition aux données journalistiques, dont l'intersection des ellipses se produit à différent degrés selon les timbres. Les voyelles de *Corpus_prep* subissent la centralisation la plus importante.

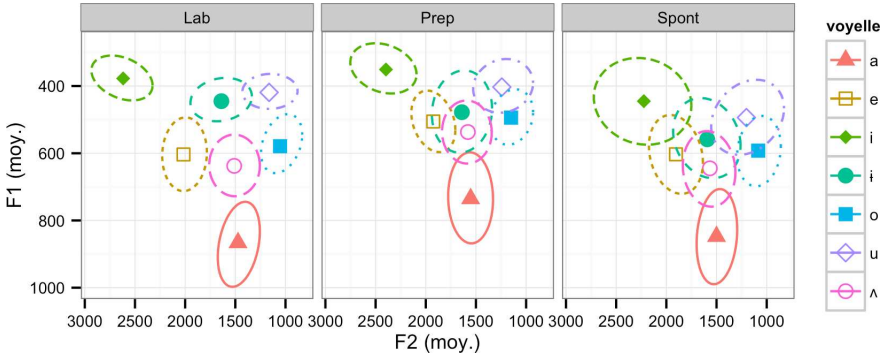


FIGURE 2 – Distribution des voyelles dans l'espace F1/F2 en fonction du style de parole (laboratoire, préparé, spontané), locuteurs féminins, en Hz.

La réduction de l'espace vocalique a comme conséquence un chevauchement des ellipses correspondant aux voyelles centrales [ʌ] et [ɨ] et postérieures sur F1. L'asymétrie sur F1 des voyelles postérieures comparées aux timbres antérieurs peut être le reflet de tendances universelles, résultant de l'anatomie du conduit vocal (De Boer, 2011). Pour ce qui est de la centralisation de [ɨ], l'effet représente une caractéristique des données journalistiques que nous ne retrouvons pas dans *Corpus_lab*. Ce chevauchement suggère une perte de contraste des deux voyelles dans la parole continue en faveur d'un timbre central. Une mesure de **la dispersion** de l'espace acoustique selon le style de parole permet de capter ces différences. La dispersion calculée comme distance moyenne entre les exemplaires des voyelles périphériques [i], [e], [u], [o], [a] et [ɨ]] (hommes et femmes) et le centroïde global est en effet plus importante pour *Corpus_lab* par rapport à *Corpus_prep* et *Corpus_spont* (Tableau 3).

	Corpus_prep	*Corpus_spont*	*Corpus_lab*
Femmes	363 (36)	359(72)	496(46)
Hommes	281(31)	286(48)	338(35)

TABLEAU 3 – Distance euclidienne au centroïde de la voyelle : valeurs moyennes et écart types par corpus (en Hz).

La **durée** oppose les données journalistiques aux données de laboratoire. Ces effets peuvent être naturellement mis en relation avec le triangle vocalique centralisé correspondant à *Corpus_prep*. La parole journalistique préparée montre ainsi des durées vocaliques plus courtes avec moins de varia-

bilité, tandis qu'aux données de laboratoire on peut associer des réalisations vocaliques plus longues (Figure 3). Par ailleurs une différence peut etre notée entre les deux types de données journalistiques *Corpus_spont* vs *Corpus_prep*, le premier présentant des durées plus longues. Le résultat, bien que faible, peut paraître quelque peu surprenant. Nous pouvons faire l'hypothèse que le cadre communicatif journalistique élicite une parole bien articulée, mais fortement compressée dans le temps : bien que moins variables, les productions vocaliques sont plus courtes et de ce fait centralisées. Par ailleurs, la parole spontanée est plus susceptible de contenir des réalisations particulièrement longues telles que les hésitations et les allongements vocaliques. L'effet de centralisation de l'espace vocalique en lien avec la diminution de la durée a été observé dans d'autres corpora comparables et semble être indépendant de la langue (Adda-Decker & Gendrot, 2010).

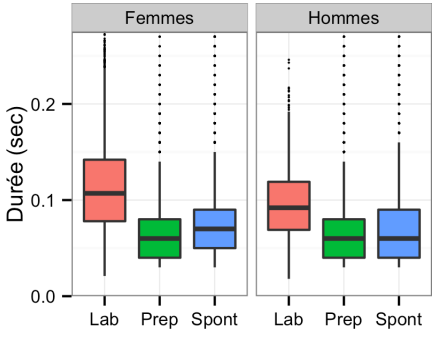


FIGURE 3 – Durée moyenne des voyelles en fonction du style de parole (laboratoire, préparé, spontané)

3.3 [ʌ] et [ɨ] : particularités acoustiques et statut phonémique

Les ellipses correspondant aux aires de dispersion de [ʌ] et [ɨ] montrent un chevauchement important sur l'axe F1 pour *Corpus_prep* et *Corpus_spont*. L'effet de réduction et la centralisation qui en est la conséquence, propres à la parole "tout venant", apportent certes des éléments d'explication, cependant tous les timbres vocaliques du système roumain ne semblent pas y réagir de la même façon. Afin d'estimer de manière plus fiable le degré de proximité des voyelles, nous avons calculé le taux de superposition des ellipses sur F1 des paires de voyelles adjacentes avec l'algorithme décrit dans (Fougeron & Audibert, 2011). Cette estimation illustrée par la figure 4 conforte l'observation empirique : une valeur positive indique une superposition des valeurs du premier formant F1. Ce taux apparaît particulièrement important notamment pour *Corpus_prep* où l'effet est renforcé par la durée réduite. Le chevauchement des ellipses de [ʌ] et [ɨ] se traduit par une centralisation forte du timbre fermé [ɨ]. Notons également que la centralisation de [ɨ] concerne toutes les productions, indépendamment du genre du locuteur et des contextes lexicaux.

L'analyse acoustique montre ainsi que les timbres des voyelles centrales [ʌ] et [ɨ] sont très proches dans la parole continue. Ce résultat corrobore l'étude de fréquence (section 3.1), une distribution complémentaire justifiant la perte de distinction des deux timbres dans la langue parlée. Le timbre central [ʌ] semble préféré, en raison probablement d'une évolution historique ([ɨ] étant à l'origine un allophone de [ʌ]). Dans l'ensemble, cette étude fournit des éléments de discussion pour le débat plus large sur la nature même du contraste phonologique, au coeur des études sur l'interface pho-

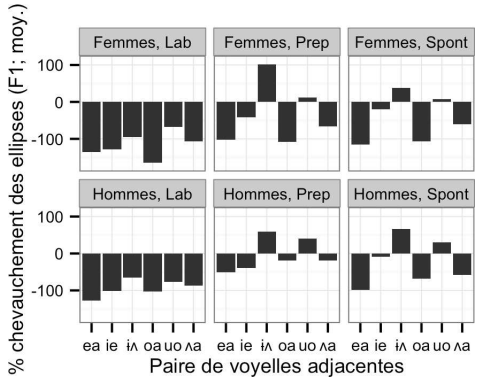


FIGURE 4 – Degré de superposition des voyelles adjacentes (une valeur positive correspond à une superposition des ellipses de deux voyelles adjacentes sur F1).

nétique/phonologie. L'exemple fourni par les voyelles du roumain appuie la description proposée par (Goldsmith, 1995) d'un contraste phonologique plutôt graduel que catégoriel. [ʌ] et [ɨ] représentent un cas de figure de voyelles "à peines contrastives" ("just barely contrastive sounds"), dont la distinction réduite en parole continue n'est pas source de perte d'intelligibilité dans la mesure où l'on peut lui associer une distribution quasi-complémentaire. Toutefois, le caractère phonémique de [ʌ] et [ɨ] perdure, grâce à des paires minimales, mais surtout à la capacité des locuteurs natifs à distinguer les deux timbres en production et perception (Renwick, 2014), soutenue probablement par un fort impact de la norme orthographique.

4 Conclusions et perspectives

Dans cet article nous avons comparé les traits acoustiques des voyelles du roumain dans la parole continue aux réalisations plus "canoniques" enregistrées en laboratoire. Les objectifs suivants ont été visés : (1) décrire les particularités acoustiques des voyelless révélées par la parole "tout venant" par rapport aux réalisations prototypiques, "de laboratoire" ; (2) estimer la relation entre traits acoustiques et contrastes phonémiques de la langue ; (3) estimer dans quelle mesure l'étude de l'oral apporte des éclairages au sujet des attributs phonémiques des voyelles centrales [ʌ] et [ɨ], dont le statut de phonèmes ou d'allophones reste controversé. Les résultats montrent des réalisations acoustiques comparables pour la parole journalistique vs de laboratoire pour l'ensemble de l'inventaire *sauf* les voyelles centrales [ʌ] et [ɨ]. Ces dernières présentent un fort chevauchement sur F1 dans la parole continue. Le résultat apporte des éclairages sur la question du contraste phonémique graduel et marginal tel que défini par (Goldsmith, 1995). Les voyelles [ʌ] et [ɨ] ont un comportement "à peine contrastif" dans la mesure où elles se trouvent en distribution quasi-complémentaire. Cette distribution mise en évidence par l'étude de fréquence dans le corpus journalistique corrobore une confusion des timbres sans perte d'intelligibilité dans la langue. Par opposition, les données de laboratoire montrent des réalisations spécifiques des deux sons, suggérant leur intégration en tant qu'éléments distincts dans les systèmes phonétique et phonémique des locuteurs natifs. Ces résultats nous encouragent à étayer l'étude des contrastes phonémiques par la prise en compte de mesures permettant l'évaluation effective de leur poids dans la langue, comme par exemple la charge fonc-

tionnelle ("functional load") (Surendran & Niyogi, 2006). Par ailleurs, une analyse plus détaillé des facteurs historiques ayant abouti à la division du timbre [ʌ] en [ʌ]/[ɨ] pourrait apporter des éclairages sur le rôle du contexte dans la préservation des attributs acoustiques des timbres respectifs. Enfin, une analyse prosodique impliquant le niveau syllabique et la prise en compte de l'accent serait sans doute bénéfique pour une meilleure description des qualités vocaliques.

Références

ADDA-DECKER M. (2006). De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. In *Actes des Journées d'Étude sur la Parole*.

ADDA-DECKER M. & GENDROT C. (2010). Influence du contexte consonantique et de la durée des voyelles sur la centralisation des voyelles orales en français. In M. EMBARKI & C. DODANE, Eds., *La Coarticulation, des Indices aux Représentations*. L'Harmattan.

AVRAM A. (1992). Despre statutul fonologic al oclusivelor palatale in limba română [sur le statut phonologique des occlusives palatales dans la langue roumaine]. *Revista de fonetică şi dialectologie [Revue de phonétique et dialectologie]*, p. 5–19.

AVRAM A. (2012). Vocalele [ă] si [î] [les voyelles [ă] si [î]]. *Studii de fonetică istorică a limbii române [Études de phonétique historique de la langue roumaine]*, p. 25–111.

BENZEGHIBA M., DE MORI R., DEROO O., DUPONT S., ERBES T., JOUVET D., FISSORE L., LAFACE P., MERTINS A., RIS C., ROSE R., TYAGI V. & WELLEKENS C. (2007). Automatic speech recognition and speech variability : a review. *Speech Communication*, **49**, 763–786.

BOERSMA P. & WEENINK D. (2001). Praat, a system for doing phonetics by computer [computer program], version 5.4.08. `http://www.fon.hum.uva.nl/praat/`.

CHITORAN I. (2002a). A perception-production study of romanian dipthongs and glide-vowel sequences. *Journal of the International Phonetic Association*, **32(2)**, 203–222.

CHITORAN I. (2002b). *The phonology of Romanian : a Constraint-Based Approach*. Studies in Generative Grammar 56. Berlin ;New York : de Gruyter Mouton.

DE BOER B. (2011). First formant difference for /i/ and /u/ : A cross-linguistic study and an explanation. *Journal of Phonetics*, **39(1)**, 110–114.

FOUGERON C. & AUDIBERT N. (2011). Testing various metrics for the description of vowel distortion in dysarthria. In *Proceedings of ICPHS*, p. 1–4.

GENDROT C. & ADDA-DECKER M. (2005). Impact of duration on f1/f2 formant values of oral vowels : an automatic analysis of large broadcas news corpora in french and german. In *Proceedings of Eurospeech-Interspeech*.

GOLDSMITH J. (1995). Phonological theory. In J. A. GOLDSMITH, Ed., *The Handbook of Phonological Theory*, p. 1–23. Cambridge, MA : Blackwell Publishers.

GRAUR A. & ROSETTI A. (1938). Esquisse d'une phonologie du roumain. *Bulletin Linguistique*, (6), 5–29.

HALL K. C. (2013). A typology of intermediate phonological relationships. *The Linguistic Review*, **30**, 215–275.

LOBANOV B. (1971). Classification of russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, (49), 606–608.

MAHALANOBIS P. C. (1971). On the generalized distance in statistics. In *Proceedings of the National Institute of Sciences (Calcutta)*, p. 49–55.

PETROVICI E. (1956). Sistemul fonematic al limbii române [système phonématique de la langue roumaine]. *Studii şi cercetări lingvistice [Etudes et recherches linguistiques]*, p. 7–14.

RENWICK M. (2014). *The Phonetics and Phonology of Contrast : The Case of the Romanian Vowel System*. Berlin, Boston : De Gruyter Mouton.

ROSETTI A. (1986). *Istoria limbii române I : de la origini pînă la începutul secolului al XVII-lea [Histoire de la langue roumaine I : des origines jusqu'au début du 17ème siècle]*. Ediţie definitivă [Edition définitive]. Bucureşti, Editura ştiinţifică şi enciclopedică.

SCOBBIE J. & STUART-SMITH J. (2008). Quasi-phonemic contrast and the fuzzy inventory : Examples from scottish english. In B. PETER AVERY & E. D. AN KEREN RICE, Eds., *Contrast in Phonology : Theory, Perception, Acquisition*, p. 87–114. Berlin : de Gruyter.

SPINU L., VOGEL I. & BUNNELL H. (2012). Palatalization in romanian-accoustic properties and perception. *Journal of Phoentics*, (40(1)), 54–66.

SURENDRAN D. & NIYOGI P. (2006). Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals. In O. NEDERGAARD THOMSEN, Ed., *Competing Models of Linguistic Change : Evolution and Beyond. In commemoration of Eugenio Coseriu (1921-2002)*, p. 43–58. Amsterdam & Philadelphia : Benjamins.

TRANDABAT D., IRIMIA E., BARBU MITITELU V., CRISTEA D. & TUFIS D. (2012). The romanian language in the digital age. In *META-NET White Paper Studies*. Springer.

VASILESCU I., VIERU B. & LAMEL L. (2014). Exploring pronunciation variants for romanian speech-to-text transcription. In *Proceedings of SLTU-2014*, p. 161–168.

La reconnaissance des mots dans la parole accentuée : Une étude en laboratoire et à l'extérieur.

Delphine Deï[1] Page Piccinini[2]
Isabelle Dautriche[1] Marieke Van Heugten[3]
Alejandrina Cristia[1]
(1) Laboratoire des Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS)
Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL Research University,
29 Rue d'Ulm, 75005, Paris, France
(2) Phonetic Lab, Department of Linguistics, UC San Diego,
9500 Gilman Drive, La Jolla, CA, United States
(3) Department of Psychology, University at Buffalo, State University of New York,
Buffalo, NY, United States
delphine.dei04@gmail.com, page.piccinini@gmail.com,
isabelle.dautriche@gmail.com, mariekev@buffalo.edu,
alecristia@gmail.com

RESUME

Des travaux récents suggèrent que les enfants et les adultes sont initialement ralentis dans leu compréhension des mots qui n'ont pas été prononcés de façon standard. Néanmoins, quand ils fon face à un interlocuteur qui à un discours accentué, ils développent rapidement des stratégie spécifiques qui leur permettent de comprendre même des prononciations atypiques. Cependant, ce résultats sont typiquement issus de recherches en laboratoire, où l'attention des participants s concentre sur une tâche unique qui leur demande peu de ressources. Afin de dépasser ce limitations, nous avons mené une expérience de reconnaissance de mots sur tablette tactile, e évaluant des enfants et des adultes, en laboratoire et dans l'environnement naturel de chaqu groupe. Nous avons constaté que des déviations de prononciation dans la parole accentué ralentissent la reconnaissance des mots, chez des enfants et adultes, tant dans le laboratoire que dan des environnements naturels.

ABSTRACT

Mispronunciations slow down word recognition: A study using touchscreens in the lab and th real world.

Recent work suggests that children and adults are initially delayed when exposed to an accente talker, but quickly recover and develop talker-specific listening patterns. This research, however, i often carried out in the lab, where participants' attention is focused on a single task and n distractions are present, and recovery is often deduced from looking patterns that require little over response. In this experiment, we overcome these limitations and assess children's and adults' wor recognition in accented speech, tested in both a laboratory and in a naturalistic setting. We foun that mispronunciations in accented speech slow down overt decisions about word recognition, to similar extent in the lab and in familiar environments, for both children and adults.

MOTS-CLES : reconnaissance de mots, parole accentuée, prononciations modifiées, acquisitio du langage, enfance.

KEYWORDS: word recognition, accented speech, mispronunciations, language acquisition toddlerhood.

1 Introduction

Bien que les accents étrangers et régionaux puissent varier à tous les niveaux phonétiques e linguistiques, nous nous adaptons rapidement aux accents auxquels nous ne sommes pas habitués Afin de mieux comprendre le système cognitif qui permet la compréhension de la parole, d nombreux travaux ont été menés sur la perception des mots dans la parole accentuée. Ces études on été conduites principalement chez des adultes, et plus récemment chez des enfants.

Plusieurs expériences en laboratoire suggèrent une flexibilité remarquable qui permet aux enfants e aux adultes de s'habituer rapidement à un nouvel accent, et à accepter des prononciations différente de celles attendues (voir Cristia et al., 2012, pour un résumé sur la perception de la parole accentué chez les enfants, jeunes adultes, et des populations plus âgées). Par exemple, Schmale, Cristia, e Seidl (2015) ont montré à des enfants de deux ans des vidéos de personnes physiquement trè différentes : un enfant, une dame âgée, une jeune fille, etc. Après cette exposition purement non linguistique, les enfants pouvaient reconnaître des mots nouvellement appris dans un accent étrange qu'ils n'avaient pas entendu auparavant. D'autres études de laboratoire coïncident avec cette vision à peines quelques minutes d'exposition suffisent pour commencer à accepter des prononciation déviantes (par ex., Maye, Aslin, & Tanenhaus, 2010; Van Heugten & Johnson, 2014).

Or, d'autres études suggèrent que les adultes et les enfants sont relativement conservateurs, en n gardant dans leur lexique que des caractéristiques typiques de leur communauté. Par exemple Girard, Floccia, et Goslin (2006) ont étudié la perception de mots prononcés dans des accent régionaux typiques de Besançon ou Toulouse auprès d'adultes résidants dans les deux régions. Il ont trouvé que la familiarité avec l'accent prédisait la performance dans une tâche de décisio lexicale, suggérant que l'expérience est cruciale pour la compréhension de la parole accentuée. Le résultats de Floccia et al. (2012) sont encore plus remarquables: les enfants de 20 mois britannique seraient *incapables* de reconnaître un mot dans la prononciation utilisée par leurs parents si cett prononciation n'est pas celle de la communauté où l'enfant grandit. Il faut cependant indiquer qu ces résultats surprenants n'ont pas été répliqués lors d'une étude plus récente menée aux Pays-Ba (Van der Feest & Johnson, 2016).

Globalement, nous voyons que la littérature ne montre pas une image monolithique et simple de l perception de la parole accentuée. Il est possible que les écarts de résultats soient, au moins e partie, dus à des différences dans la complexité de la tâche, dans la façon dans laquelle elle a ét conçue, et dans la prévalence des accents dans le lieu où elle a été menée. Une deuxième limitatio des études antérieures est qu'elles ont été conduites dans des lieux inhabituels tels que le laboratoires, bien loin des conditions du monde réel. En effet en dehors du laboratoire nous somme bien souvent entourés d'autres individus, et probablement plus ouverts à la possibilité de rencontre des accents et individus variés. Cette différence pourrait se révéler capitale car nos modèles de l perception de la parole sont donc basés sur des expériences peu écologiques qui pourraient ne pa être généralisables à la perception de la parole comme elle se déroule dans la vie de tous les jours Or, du moins à notre connaissance, la possibilité que le contexte expérimental joue un rôle dans l flexibilité face à des prononciations peu standards n'a pas été étudiée.

L'expérience que nous proposons aborde ces limitations et compare en laboratoire et dans un lieu quotidien la reconnaissance de mots accentués. Des travaux précédents ont montré que la perception de la parole accentuée est facilitée par des indices extérieurs, tels que la présence d'une photo révélant l'identité du locuteur (Johnson, Strand, & D'Imperio, 1999) ou même la présence d'un mo ou d'un signe régional (Hay & Drager, 2010). Notre étude ne cherche pas à étudier ces biais, mai plutôt à déterminer si les participants qui entendent une parole accentuée sont plus flexibles, c'est à dire, acceptent des prononciations déviantes plus facilement, quand ils sont testés dans un environnement quotidien, par rapport au laboratoire. Nous avons donc recruté des adultes et de jeunes enfants, car l'impact de l'environnement pourrait être plus marqué chez ces derniers.

Tous les participants ont réalisé une tâche de choix forcé avec des stimuli enregistrés auprès d'une personne ayant un accent étranger (anglais britannique, un accent commun à Paris). Les participants avaient en main une tablette tactile où 2 images à la fois étaient visibles (par ex. une maison, une chaussure) et où un personnage animé leur demandait d'appuyer sur l'une d'elle (ex "Touche la maison"). Les objets visibles n'avaient aucun recoupement phonologique de sorte que le choix était toujours évident. Or, dans certains essais, le mot n'était pas prononcé de la façon attendue (par ex. "***b***ateau" devenait "***p***ateau" avec l'accentuation anglaise). Nous voudrions savoir si les enfants et les adultes sont moins affectés par ces déviations dans un environnement quotidien par rapport à la même tâche en laboratoire. Puisque notre recrutement se passait en partie dans des lieux publics, il a été impossible de faire des groupes *a priori* sur la base de l'expérience linguistique des participants. Néanmoins, un questionnaire nous a permis de contrôler ces caractéristiques post hoc.

2 Méthode

2.1 Participants

Nous présentons les caractéristiques des participants séparés par groupe d'âge (adultes, enfants) et du lieu où le test s'est déroulé (laboratoire, quotidien), comme illustré dans la Table 1. Le premier groupe d'enfants a été testé au laboratoire bébé du LSCP (groupe « Enfant-Labo »). Le deuxième groupe a été testé à la crèche l'Arbre-Sec (Paris 1er) pour les plus jeunes et à l'école François Coppée (Paris 15e) en petite section pour les plus grands (groupe « Enfant-Quotidien »). De la même façon, nous avons testé une partie des adultes en cabine de test au laboratoire (« Adulte Labo »), et d'autres à la bibliothèque universitaire de l'UPMC (« Adulte-Quotidien »). Dans tous les cas, les participants ou leurs représentants légaux ont donné leur consentement par écrit.

À l'intérieur de ces quatre groupes, des participants ont du être exclus en suivant un critère d'exposition au français non-accentué minimum de 80% du temps, ainsi que 7 critères d'exclusion fixés avant l'inspection des résultats : 1) Le questionnaire d'expérience linguistique n'a pas été rempli ; 2) Le français est entendu moins de 50% du temps ; 3) Réponses incorrectes aux deux essais de familiarisation ; 4) Plus de 25% de réponses incorrectes aux essais correctes pour un francophone ; 5) Avoir répondu à moins de 9 essais test sur les 12 ; 6) Avoir subi une ou plusieurs otites pendant la semaine qui précède le test ; 7) Avoir été diagnostiqué avec un trouble du développement (hors prématurité).

	Total → Inclus	**Âge (écart type)**
Enfant-Labo	19 (9) → 9	42,25 (± 8,13) mois
Enfant-Quotidien	48 (13) → 18	36,80 (± 7,45) mois
Adulte-Labo	15 (10) → 8	23,17 (± 4,37) ans
Adulte-Quotidien	28 (15) → 11	22,77 (± 3,78) ans

TABLE 1 : Informations démographiques des groupes testés. La colonne "Total" reporte le nombre total d'individus testés pour chaque groupe (et le nombre de femme parmi eux). "Inclus" indique le nombre d'individus inclus dans nos analyses. "Âge" indique l'âge moyen (et l'écart type) pour les individus qui ont pu être inclus

2.2 Matériel et procédure

Nous avons mesuré le temps de réaction avec une tâche de choix forcé implémentée sur une tablett tactile numérique iPad®. Pour rendre l'activité plus ludique et attractive pour les enfants, les énoncé étaient adaptés à leur âge et formulés par un personnage animé dans un décor coloré (voir Figure 1) Dans chaque essai une paire d'objets était visible et l'un des deux était nommé par le personnag dans la phrase « Touche le __ ». Les essais n'étaient pas répétés, même en cas de répons incorrecte. En revanche, le personnage sur l'écran renvoyait un feedback positif en cas de répons correcte (« Super ! », en sautant de joie) et négatif en cas de réponse incorrecte (« Non, c'est pa celui la », tristement). L'expérimentateur donnait aux enfants des encouragements neutres à la tâch (« tu te débrouilles très bien ») afin de maintenir son intérêt.

FIGURE 1: Capture d'écran de l'iPad® pendant le jeu pour un essai avec le couple banane/chaussette

Il y avait au total 14 essais. Les deux premiers ont servi à familiariser les participants avec la tâche et au fait que le personnage avait un accent étranger. Uniquement durant cette phase les individu testés pouvaient recevoir du feedback de l'expérimentateur sur leur performance. Nous avons don utilisé des stimuli neutres à notre question de recherche pour cette première phase, les réponse n'étaient prises en compte que dans les critères d'exclusion définis plus haut.

Les 12 essais restants étaient répartis en 3 conditions différentes : **Prononciation française (PF)**, l mot est prononcé avec la forme phonologique attendue avec une prononciation française standard **Prononciation anglaise (PA)**, le mot est prononcé avec la forme phonologique attendue avec un

prononciation accentuée anglaise, par exemple comme vu plus haut: "***b***ateau" devient "***p***ateau" ; e **Prononciation composée (PC)**, le mot est prononcé d'une façon inattendue pour l'accentuatio française et anglaise, par exemple le mot "***b***anane" devient "***v***anane". Les mots test utilisés dans ce trois conditions avaient soit des voyelles nasales, soit des voyelles antérieures arrondies, soit de consonnes occlusives voisées (voir section Stimuli). Ces associations étaient contrebalancées à travers les participants ; par exemple, pour un participant donné tous les mots ayant des voyelle nasales étaient prononcés correctement et donc représentaient la condition PF ; les voyelles de tou les mots avec des voyelles antérieures arrondies changeaient de hauteur, représentant ainsi l condition PC ; et toutes les consonnes occlusives voisées des mots restants étaient dévoisées pou représenter la condition PA.

2.3 Stimuli

Les stimuli ont été produits par une femme qui a pour langue maternelle l'anglais britannique et qu parle français couramment. Nous les avons traité sous PRAAT (Boersma & Weenink, 2015), en nous assurant qu'ils ne différaient ni en intensité ni en durée à travers les 3 conditions (en moyenne).

Les stimuli sélectionnés étaient des noms connus par plus de 60% d'enfants d'après un test d vocabulaire donné à 81 enfants de 18 mois lors d'une étude précédente. Nous avons ensuit sélectionné 14 paires de mots [cibles + distracteurs], 2 pour les essais de familiarisation et 12 pou les essais test. Nous avons exclu les mots qui contenaient plus d'un type de contraste; comme pa exemple le mot "ballon" qui possède une consonne occlusive voisée, /b/, et une voyelle nasale, /ɔ̃/ Par ailleurs, nous nous sommes assurés que les prononciations modifiées pour les conditions PA e PC, de tous les mots cibles, ne formaient pas de vrais mots. Les mots prononcés pour la phase d familiarisation et les phrases de transition (feedback par exemple) ne contenaient pas de contraste affectés par l'accent anglais.

Les distracteurs étaient choisis pour avoir le même genre grammatical, nombre de syllabes, et l même animéité que la cible. De plus, les mots cibles et leurs distracteurs appariés étaient clairemen phonologiquement différents, ainsi, même si la prononciation était différente de celle attendue, l seule réponse possible restait l'image cible. Les paires cible-distracteur choisies étaient les suivante : pour les voyelles nasales *poisson-lapin*, *maison-chaussure*, *manteau-gâteau*, *compote-poussette* pour les voyelles arrondies *lunettes-pantoufles*, *voiture-poubelle*, *fleur-bouche*, *yeux-chiens* ; pou les consonnes voisées *balle-main*, *bateau-cochon*, *bouteille-cuillière*, *banane-chaussette*.

2.4 Analyses et prédictions

Notre question de recherche principale concerne la flexibilité en perception face à de prononciations non-standards à travers différent environnements de test. Le temps de réponse es défini comme le temps écoulé entre la fin de l'énoncé (ex. Touche la maison) et le moment du touché sur la tablette. Afin d'obtenir une meilleure appréciation des résultats en éliminant le variations individuelles, les différences d'âge, et les variations d'ordre général sur le temps d réponse, nous avons choisi de calculer les ratios des temps de réponse sous R (R Core Developmen Team, 2008) tel que:

$$([PA] / [PA] + [PF]) - 0,5 \text{ et } ([PC] / [PC] + [PF]) - 0,5$$

Où [PA] correspond à la médiane des temps de réponse pour la prononciation anglaise, [PF] à la médiane des temps de réponse pour la prononciation française et [PC] celle pour la prononciation composée.

Sur la base des travaux antérieurs, et compte tenu du fait que le test était très court, nous nous attendons à ce que les ratios soient supérieurs à 0 au laboratoire, indiquant que les participants sont ralentis par les prononciations inattendues (PA, PC) par rapport à la prononciation attendue (PF). Ceci répliquerait des résultats antérieurs montrant que les enfants et adultes sont sensibles à la prononciation d'un mot même lorsqu'il est prononcé avec un accent étranger. Notre question principale porte sur la flexibilité en perception dans un environnement quotidien par rapport au laboratoire. Si les participants sont plus flexibles dans un environnement quotidien, alors leurs ratios devraient être significativement plus bas dans cet environnement par rapport au laboratoire. Par contre, si le contexte ne fait aucune différence, le ratio devrait être aussi élevé (et supérieur à 0) pour chaque groupe d'âge dans chaque milieu. Pour comparer les ratios, nous utiliserons un test non paramétrique de Wilcoxon entre les deux sous-groupes (environnement quotidien versus laboratoire) à l'intérieur de chaque groupe d'âge.

3 Résultats

Par soucis de simplicité, nous présentons seulement les ratios pour la condition PC, mais le pattern de résultats dont nous concluons est le même pour la condition PA. Toutes nos données et nos analyses sont disponibles sur la plateforme Open Science Framework (OSF), et directement accessibles depuis https://osf.io/pnhc3/.

Tout d'abord, nous observons que les enfants sont en général plus lents que les adultes, et que les enfants testés au laboratoire sont plus lents que ceux testés dans leur environnement quotidien. Enfants-Labo(EL)=3,93s, EQ=3,47s, AL=2,26s et Adultes-Quotidien (AQ)=2,35s en moyenne. Tous les ratios sont significativement différents de 0 pour tous les groupes (voir Figure 2). Cela signifie que, comme prédit, les temps de réponse pendant des essais où les mots sont prononcés de manière inattendue (prononciation anglaise ou composée) sont supérieurs à ceux des mots prononcés en français standard. Enfin, cet effet n'est pas modulé par le lieu d'expérimentation.

Des analyses supplémentaires ont été faites mais n'ont pas être présentées ici par manque d'espace. Elles peuvent être reproduites en utilisant les données et le code disponible sur notre site OSF. Notamment, nous avons évalué la possibilité qu'un petit nombre de participants ayant une expérience avec d'autres langues ou accents supérieure à 20% de leur temps soient plus flexibles avec les prononciations inattendues. Contrairement à cette hypothèse, nous ne trouvons pas de différence entre les deux groupes de participants (avec et sans expérience). Nous ne trouvons pas non plus de différences marquées parmi les enfants en fonction de leur exposition préalable aux écrans tactiles, ni à travers les différents types de changements (voyelles versus consonnes).

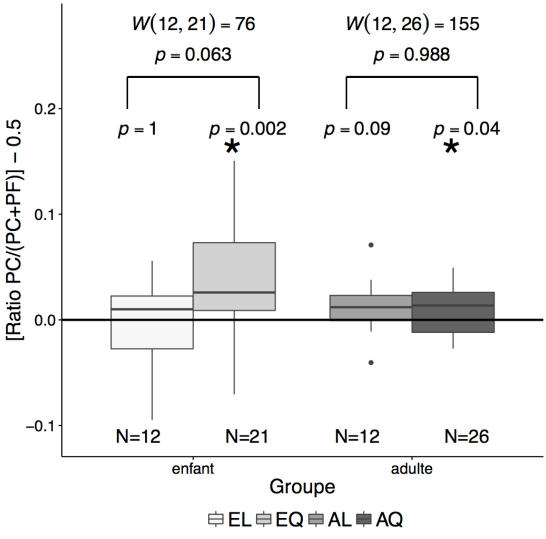


FIGURE 2: Graphique en boîte à moustache représentant le ratio des médianes des temps de réponse pour les essais des mots modifiés par la prononciation composée, dans les différentes conditions d'expérimentation (en laboratoire ou à l'extérieur) pour les deux groupes testés (adultes et enfants).

4 Discussion

Dès le plus jeune âge, la perception et la compréhension du langage est un enjeu important, car les enfants doivent rapidement apprendre une langue et interagir avec des personnes diverses présentes dans leur environnement. Nous avons ainsi étudié la compréhension des mots chez des enfants âgés de 24 à 48 mois, en parallèle avec des adultes.

Nous avons répliqué des études antérieures montrant que les prononciations non-standards ont un coût cognitif : les adultes et les enfants sont ralentis quand un mot n'est pas prononcé de façon standard, et ce même en présence de signes que l'interlocuteur peut ne pas être précis, puisqu'il a un accent étranger. Ce qui constitue la principale nouveauté de notre étude est que nous avons mesuré ce coût non seulement en laboratoire, mais aussi dans des environnements quotidiens pour nos participants. Nos résultats montrent que la perception de la parole accentuée n'est pas affectée différemment lorsqu'un étudiant est testé à la bibliothèque ou au laboratoire ; ni lorsqu'un enfant est testé à la crèche ou à l'école par rapport à une cabine insonorisée et isolée. Ceci pourrait indiquer que, bien que le contexte environnemental puisse biaiser la perception de façon globale (par ex. Hay & Drager, 2010), il ne change pas les mécanismes basiques de compréhension de mots.

Bien sur, il est possible qu'une différence existe à travers les environnements de test, mais elle serait trop petite pour être détectée avec la procédure employée ici. En effet, nous avons utilisé un plan expérimental où le facteur contexte varie à travers les groupes, ce qui résulte à moins de puissance statistique qu'un dessin intra-sujet. Bien que des recherches ultérieures pourraient corriger cette

limitation, nous attirons l'attention sur le fait que la tendance n'est pas vers plus de tolérance pou les prononciations inattendues dans un environnement quotidien, mais l'inverse.

Cette étude, comme d'autres, pourraient profiter de la validation d'une technique encore très peu utilisée en recherche fondamentale, le test sur tablette tactile, dont nous avons fait usage. Le fait qu nous avons pu détecter le ralentissement en perception induit par une prononciation inattendue même dans un environnement peu contrôlé, suggère que cette technique pourrait nous permettre d rendre nos protocoles plus facilement transportables, et d'utiliser un support stable dans des lieu pas spécialement adaptés aux tests scientifiques (crèches, écoles, etc.)

Remerciements

Ce travail a été possible grâce au support économique et institutionnel de ANR-10- LABX-008 IEC (à travers de leur fonds d'aide aux chercheurs arrivants), ANR-10-IDEX-0001-02 PSL* (pa ses Action Incitatives), et ANR-14-CE30-0003 MechELex.

Références

BOERSMA, P., WEENINK, D. (2015). Praat, a system for doing phonetics by computer. Amsterdam.

CRISTIA, A., SEIDL, A., VAUGHN, C., SCHMALE, R., BRADLOW, A., FLOCCIA, C. (2012). Linguistic processing of accented speech across the lifespan. *Frontiers in psychology*, *3*, 479.

FLOCCIA, C., DELL LUCHE, C., DURRANT, S., BUTLER, J., GOSLIN, J. (2012). Parents or community Where do 20-month-olds exposed to two accents acquire their representations of words? *Cognitio* 124, 95-100.

GIRARD, F., FLOCCIA, C., GOSLIN, J. (2006). Familiarité aux accents régionaux et identification d mots. *Actes des JEP 2006*, 449-452.

HAY, J., DRAGER, K. (2010). Stuffed toys and speech perception. Linguistics, 48(4), 865-892.

JOHNSON, K., STRAND, E. A., D'IMPERIO, M. (1999). Auditory–visual integration of talker gender i vowel perception. Journal of Phonetics, 27(4), 359-384.

MAYE, J., ASLIN, R. N., TANENHAUS, M. K. (2010). The Weckud Wetch of the Wast: Lexica adaptation to a novel accent. *Cognitive Science* 33, 543-562.

R DEVELOPMENT CORE TEAM. (2008). R: A language and environment for statistical computing. F foundation for Statistical Computing. Vienna, Austria.

SCHMALE, R., SEIDL, A., CRISTIA, A. (2015). Mechanisms underlying accent accommodation i early word learning: Evidence for general expansion. *Developmental science*, *18*(4), 664-670.

VAN DER FEEST, S. V., JOHNSON, E. K. (2015). Input-driven differences in toddlers' perception of disappearing phonological contrast. *Language Acquisition*, 1-23.

VAN HEUGTEN, M., JOHNSON, E. K. (2014). Learning to contend with accents in infancy: Benefits o brief speaker exposure. *Journal of Experimental Psychology: General*, *143*(1), 340.

Répartition des phonèmes réduits en parole conversationnelle. Approche quantitative par extraction automatique

Christine Meunier[1] & Brigitte Bigi[1]

(1) Laboratoire Parole et Langage, Aix Marseille Université, CNRS
5 avenue Pasteur, 13100 Aix-en-Provence, France
christine.meunier@lpl-aix.fr, brigitte.bigi@lpl-aix.fr

RESUME

Cette étude vise à mieux comprendre la répartition des réductions phonétiques présentes dans la production de parole. Nous avons sélectionné lôensemble des phonèmes les plus courts (30ms) à partir de lôalignement dôun corpus de parole conversationnelle. Cette version contenant uniquement les phonèmes courts (V1) est comparée à la version contenant lôalignement de tous les phonèmes du corpus (V0). Les deux versions sont mises en relation avec lôannotation des mots et de leur catégorie syntaxique. Les résultats montrent que les liquides, les glissantes et les voyelles fermées sont plus représentées dans V1 que dans V0. Par ailleurs, la nature et la catégorie syntaxique des mots modulent la distribution des phonèmes en V1. Ainsi, la nature instable du /l/, ainsi que sa présence dans de très nombreux pronoms et déterminants, en fait le phonème le plus marqué par la réduction. Enfin, la fréquence des mots semble montrer des effets contradictoires.

ABSTRACT

The distribution of reduced phoneme in conversational speech. A quantitative approach by automatic extraction.
This study aims to provide a better understanding of the distribution of phonetic reduction in speech production. Shortest phonemes (30ms) have been selected from the alignment of a conversational speech corpus. This version, containing all (and only) the shortest phonemes (V1), is compared to the basic version (V0) containing the all phonemes. The annotation of lexical and morpho-syntactic categories (CMS) is added in both versions. Results show that liquids, glides and closed vowels are proportionally more present in V1 than in V0. Moreover, words and their CMS modulate phoneme distribution in V1. Thus, /l/ is an unstable phoneme and itôs also present in very frequent function words such as pronouns or determinants. This makes it the best candidate for reduction. Finally, word frequency shows contradictory effects.

MOTS-CLES : Réduction phonétique, parole spontanée, alignement, fréquence lexicale
KEYWORDS: Phonetic reduction, spontaneous speech, alignment, lexical frequency.

1 Introduction

Notre objectif, dans ce travail, est dôaborder la question des réductions phonétiques via une approche ascendante et quantitative. Pour ce faire, nous exploitons les données de lôalignement automatique sur un corpus de parole conversationnelle en faisant lôhypothèse que lôanalyse des

segments de durée minimale (30ms) peut fournir des informations intéressantes quant au contexte linguistique (lexical et infra-lexical) des réductions phonétiques.

La réduction phonétique peut être définie concrètement comme une production sous-spécifiée des segments phonétiques caractérisée par un éloignement des cibles prototypiques. Cette sous-spécification peut se manifester par des omissions de phonèmes ou encore des changements dans les caractéristiques acoustico-articulatoires des phonèmes. De nombreux travaux ont pu mettre en évidence les modifications phonétiques caractéristiques de la réduction, comme la réduction vocalique liée au débit (Gendrot et Adda-Decker, 2007 ; Meunier et Espesser, 2011) ou encore lôassimilation de voisement dans les séquences de consonnes en parole spontanée (Duez, 1995 ; Hallé & Adda-Decker, 2007). Mais quelles que soient les caractéristiques de la réduction, elle est le plus souvent concomitante avec lôaugmentation du débit (Pluymaekers et al. 2005, Gendrot et Adda-Decker, 2007).

Mais décrire les caractéristiques de la réduction ne peut être une fin en soi. Il est également nécessaire dôapporter un éclairage sur le fonctionnement de la réduction et donc dôétablir un lien entre les zones réduites et dôautres paramètres de la communication linguistiques ou extralinguistiques. Les locuteurs, dans leur production courante, adaptent leurs commandes motrices en fonction des contraintes de la communication (Lindblom, 1990). La réduction phonétique peut donc être considérée comme la **latitude** dont dispose le locuteur pour **optimiser** sa production. Cette optimisation est assujettie aux compétences linguistiques de chaque locuteur (les connaissances quôil a de sa propre langue). On peut donc estimer que la façon dont les locuteurs utilisent cette latitude nous éclaire sur **lôusage** des compétences linguistiques. Plusieurs travaux se sont focalisés sur les propriétés lexicales des mots, comme la fréquence (Jurafsky et al. 2001, Pluymaekers et al. 2005), la catégorie dôusage (mots fonction/mots de contenu, Johnson, 2004), ou encore la nature de certaines formes fréquentes (omission de /E/ dans la séquence « cô était » en français, Torreira & Ernestus, 2011). Dans lôensemble de ces travaux, un lien est établi entre la réduction phonétique et la fréquence, lôusage (mots fonction) ou la prédictibilité des mots. Néanmoins, dans la grande majorité de ces travaux, lôapproche utilisée est descendante alors que nous visons ici une approche ascendante où lôextraction automatique des segments courts est mise en relation avec des propriétés supra-phonétiques. Les approches quantitatives issues du traitement automatique ont apporté un éclairage considérable concernant la variation et la réduction phonétique dans des grands corpus de parole (Adda-Decker et al., 2008). Elles apportent une information complémentaire visant à traiter la globalité de la réduction. Identifier et localiser les segments extra-courts dans un grand corpus de parole spontanée nous semble donc une bonne entrée pour comprendre la répartition et le fonctionnement de la réduction dans la parole.

Notre objectif dans ce travail est donc de mieux comprendre la nature des segments extra-courts (i.e. réduits). A ce stade, nous abordons trois questions : 1/ la réduction phonétique affecte-t-elle indifféremment tous les phonèmes, ou bien certains phonèmes sont-ils plus aptes à réduire ? 2/ les segments réduits se concentrent-ils plus volontiers sur certains mots ou catégories de mots ? 3/ la fréquence des mots joue-t-elle un rôle dans la réduction des phonèmes ?

2 Méthode

Nous avons choisi dôobserver la réduction via une extraction automatique des segments phonétiques les plus courts. Avant cette étape, un repérage manuel des zones réduites avait été effectué. Ce travail nous avait permis dôétablir un lien entre les zones de réduction phonétique et

lôaccumulation de segments de durée minimale (Meunier, 2012). Lôidentification de ces segments se présente ainsi comme un bon indicateur de la présence de réduction.

Les données ont été extraites du *Corpus of Interactional Data* (CID, Bertrand, 2008). Il sôagit dôun enregistrement audio-vidéo de dialogues spontanés entre des locuteurs français natifs (16 conversations dôune heure chacune entre deux locuteurs, soit 16 locuteurs, 10 femmes et 6 hommes). Nos descriptions se basent sur un style de parole spontanée et relâchée (conversations familières). Lôextraction des données a été effectuée sur une version du corpus phonétisée à partir dôune Transcription Orthographique Enrichie, puis alignée avec SPPAS (Bigi, 2015) et taggée avec Marsatag (http://sldr.org/sldr000841). Ces annotations du CID sont publiquement disponibles http://sldr.org/sldr000720). Cette version complète de lôalignement phonétique constitue la version de référence (**V0**). Une seconde version (**V1**) a été produite en extrayant tous les segments de 30ms (correspondant à la trame minimale de lôaligneur). Les analyses présentées plus loin ont été effectuées grâce à une comparaison entre la version originale de lôalignement (V0) et la version comprenant uniquement les segments de 30ms (V1). Cette comparaison permet dôévaluer la proportion ainsi que la répartition des phonèmes réduits (i.e. de 30ms).

Notre objectif est de mettre en relation ces segments extra-courts avec dôautres strates linguistiques intégrées dans les annotations du CID. Pour les deux versions (V0 et V1) nous disposons ainsi dôune annotation sur trois niveaux comprenant : 1/ lôalignement phonétique, 2/ lôannotation en mots (*Tokens*) et 3/ lôannotation en Catégories Morpho-Syntaxiques (CMS) (figure 1). Par soucis de clarté, les unités phonétiques seront transcrites en code SAMPA[1] dans lôensemble du texte et des figures (https://www.phon.ucl.ac.uk/home/sampa/french.htm).

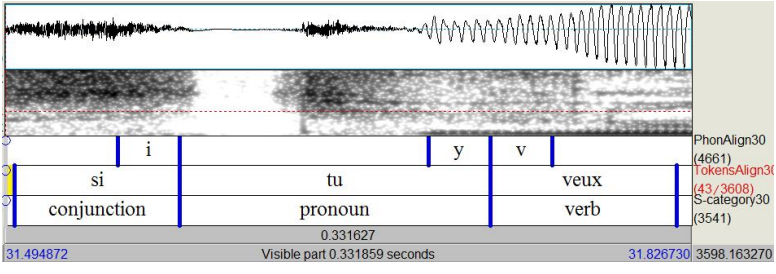


FIGURE 1: Les trois niveaux dôannotation produits dans la V1 (segments de 30ms, Tokens correspondants, Catégorie Morpho-Syntaxique correspondantes).

Nous avons bien conscience que la durée des phonèmes dans un corpus aligné automatiquement est liée aux caractéristiques, voire aux biais de lôaligneur. Ces segments de 30ms peuvent donc contenir des erreurs mais aussi des segments non réalisés (car transcrits). Toutefois, les omissions ne sont pas un problème pour ce travail : un phonème transcrit mais non réalisé est, de toute façon considéré comme réduit. Ça nôest donc pas la durée des phonèmes elle-même qui nous intéresse ici, mais lô*indication* quôun phonème est réduit. Il est par ailleurs possible que lôaligneur sous-estime ou, à lôinverse, sur-estime la durée de certains phonèmes. Pour cette raison, les frontières produites par lôaligneur ont été comparées avec une segmentation manuelle faite par deux experts. La différence entre les deux annotations ne montre pas dôécart majeur pouvant avoir un impact sur la présente étude.

[1] Notons que la version fournie par lôaligneur ne fait pas de distinction à lôintérieur des macro-classes de voyelles. Ainsi A regroupe /a/ et /ɑ/, E regroupe /e/ et /ɛ/, /O/ regroupe /o/ et /ɔ/ et & regroupe /ø/, /œ/ et le schwa /ə/.

3 Résultats

3.1 Réduction et propriétés phonétiques

Les effectifs de V0 et V1 nous indique que les segments de 30ms (V1=48129 occurrences) représentent 16,4% de la totalité des phonèmes du corpus (V0=293591 occurrences). Tous les phonèmes sont représentés dans V1, même sôils ne le sont pas tous dans les mêmes proportions. La répartition des phonèmes fréquents ou rares dans des grands corpus de parole, quôils soient oraux ou écrits, sont en général assez stables. On retrouve toujours parmi les phonèmes les plus fréquents, les voyelles /A/, /E/, /i/, /&/ et les consonnes /R/, /s/, /t/, /l/ (New et al., 2001). Dans nos données, la hiérarchie habituelle est globalement respectée (Table 1). Les phonèmes /E/, /A/, /R/ restent en tête des phonèmes les plus fréquents aussi bien en V0 quôen V1. Autrement dit, tous les phonèmes fréquents présentent un nombre conséquent de réalisations réduites.

	E	A	R	s	t	l	i	k	&	m	p	d
V0	38592	28235	17514	17261	16508	15458	14036	13660	12186	11980	11549	10989
V1	**5674**	**3342**	**3603**	735	1639	**5987**	**3078**	1261	**3155**	1218	926	1432
(V1/V0)*100	15	12	**21**	4	10	**39**	**22**	9	**26**	10	8	13

TABLE 1: nombre dôoccurrences des 12 phonèmes les plus fréquents dans V0 (ligne 1, par ordre décroissant) ; nombre correspondant dans V1 (ligne 2); pourcentage de V1 dans V0 (ligne 3). En gras et rouge, les 6 phonèmes les plus fréquents dans V1 (/l/, le plus fréquent, est encadré).

On trouve toutefois un certain nombre de disparités entre les deux versions. Ces disparités nous informent sur les phonèmes qui sont sous- ou sur-représentés en V1 (figure 2, respectivement, à gauche et à droite). Notamment, on observe en V1 une sur-représentation des liquides (particulièrement /l/, mais aussi /R/ dans une moindre mesure), des glissantes (particulièrement /w/, mais aussi /j/ et /H/), des voyelles fermées et du /&/ (regroupant /@/, /2/ et /9/) (figure 2, à droite). La sur-représentation en V1 indique que les phonèmes ont une tendance forte à la réduction. En revanche, les occlusives et fricatives sourdes (essentiellement /s/ et /t/) ainsi que les nasales (voyelles et consonnes) sont plutôt sous-représentées en V1 (figure 2, à gauche).

FIGURE 2: Différence (V0-V1) des pourcentages dôoccurrence des phonèmes pour chaque jeu de données (V0 et V1). Ex : la valeur de /s/ représente la différence entre son pourcentage dôoccurrences dans V0 (5.7%) et celui de V1 (1.5%), soit 4,2. Les phonèmes dont la valeur apparaît en négatif sont donc sur-représentés en V1 par rapport à V0.

Nous avons par ailleurs calculé le taux de réduction pour chaque phonème (% de V1 dans V0). Ce taux confirme une réduction importante pour les glissantes et le /l/. La glissante /H/, peu fréquente dans le corpus, est le phonème dont le taux de réduction est le plus important (80% de ses occurrences sont de 30ms). Il en va de même, pour /l/ (en revanche bien plus fréquent) dont 39% des occurrences font 30ms. Cela en fait le phonème le plus réduit dans la mesure où il est le plus fréquent en V1. Pour comparaison, /E/, la voyelle la plus fréquente aussi bien en V0 quôen V1, ne montre un taux de réduction que de 15%, ce qui suppose quôune grande partie de ses réalisations ne sont pas réduites (voir table 1).

Nous allons voir cependant, quôen dépit des propriétés acoustiques et articulatoires des segments phonétiques, les propriétés lexicales (catégorie et fréquence) dans lesquels ces segments sont produits modulent considérablement leur apparition en V1.

3.2 Réduction et propriétés lexicales

Le contexte lexical est abordé ici selon plusieurs facteurs : mots, catégories morpho-lexicales et fréquence des mots. Ces facteurs sont, encore une fois, regardés à la lumière des différences entre V0 et V1. Pour V1, lôoccurrence dôun mot correspond à la présence dôau moins un phonème de 30ms dans ce mot.

Les mots les plus fréquents sont répartis différemment entre V0 et V1 (table 2). Ainsi, le pronom « cô », très fréquent dans le corpus, apparaît très peu dans V1. La nature phonétique de /s/, peu enclin à réduire, conditionne probablement cette sous-représentation. Il en va de même pour les mots « et », « pas », « ça », « ouais » « est », « à ». Notons que tous ces mots contiennent les phonèmes /E/, /A/, /p/, /s/ qui sont des phonèmes sous-représentés en V1 (figure 2).

	est	**c'**	**ouais**	**et**	**tu**	**de**	**pas**	**ça**	**le**	**il**	**je**	**mais**	**que**	**a**
V0	3423	3293	3184	2914	2343	2281	2142	2071	1915	1751	1739	1619	1600	1455
V1	**799**	134	**631**	171	**916**	**610**	195	273	**796**	**688**	454	324	532	190
(V1/V0)*100	23	4	20	6	**39**	27	9	13	**42**	**39**	**26**	20	**33**	13

TABLE 2: nombre dôoccurrences des 14 mots les plus fréquents dans V0 (ligne 1, par ordre décroissant) ; nombre correspondant dans V1 (ligne 2) ; pourcentage de V1 dans V0 (ligne 3). En gras et rouge, les 6 mots les plus fréquents dans V1 (« le », le plus fréquent, est encadré).

A lôinverse, les mots « tu », « quoi », « le », « il », « la », « de » sont les plus fréquents de V1 et contiennent des phonèmes sur-représentés en V1 (/l/, /&/, /w/, /i/). On notera toutefois que les écarts de proportion entre V0 et V1 sont relativement faibles.

La présence des **catégories morpho-syntaxiques** (CMS) diffèrent peu entre V0 et V1. La forte prédominance des verbes, pronoms et noms est préservée en V1. Pour mettre en évidence le lien entre réduction phonétique et CMS, nous avons choisi 6 phonèmes intéressants pour cette analyse (car fréquents et/ou montrant une différence importante entre V0 et V1). Pour ces 6 phonèmes, nous avons regardé leur distribution dans 6 CMS (représentatives des mots de contenu et mots fonction). La figure 3 montre cette distribution dans V0 (à gauche) et V1 (à droite). On note pour la consonne /l/ une très forte augmentation de sa présence en V1, et cela pour lôensemble des 6 CMS. Cela suppose que la réduction de /l/ est homogène dans les 6 CMS. En revanche, pour /R/ et /w/ lôaugmentation porte essentiellement sur les CMS dans lesquelles ces deux consonnes sont déjà présentes en V0 : on ne note pas dôaugmentation des réductions dans les mots fonctions

(excepté pour les pronoms dans lesquels /w/ est présent). A l'opposé, la présence des voyelles /E/ et /A/ baisse globalement en V1, excepté pour la catégorie des verbes. Il est donc probable que la flexion verbale joue ici un rôle important. Enfin, on notera la diminution très importante de /s/ dans V1 qui est cohérente avec l'observation faite sur la figure 2 (notons également que /l/ et /s/ ont des comportements totalement inversés). La relation entre réduction phonétique et CMS semble donc très complexe et dépend à la fois des facteurs phonétiques et morphologiques.

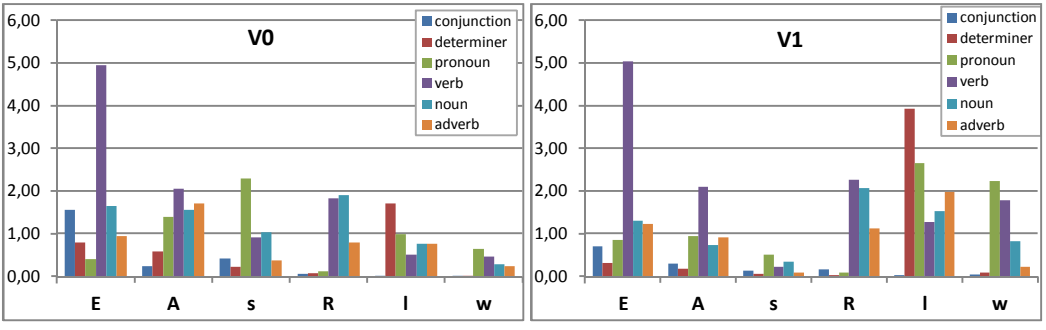


FIGURE 3: Pourcentage d'une sélection de 6 phonèmes dans V0 (à gauche) et V1 (à droite) en fonction de leur présence dans 6 CMS.

Nous avons par ailleurs cherché à évaluer si la présence de phonèmes réduits est identique pour les **mots rares et les mots fréquents**. Nous avons ainsi comparé la répartition des phonèmes de 30ms (V1) dans deux types de mots : 1/ les mots BF (basse fréquence) regroupant tous les mots qui apparaissent au plus 2 fois dans le corpus pour V1 et au plus 5 fois pour V0 ; 2/ les mots HF (haute fréquence) regroupant tous les mots qui apparaissent au moins 250 fois V1 et au moins 1000 fois pour V0. Nous avons en effet différencié ces critères pour V1 et V0 de façon à ce que les effectifs soient proportionnellement comparables (table 3).

	HF	BF
V0	63312 occurrences (22% des mots)	55216 occurrences (19% des mots)
V1	9342 occurrences (19% des mots)	8788 occurrences (18% des mots)

TABLE 3: Effectifs et proportion des mots HF et BF dans V0 et V1.

Sur la figure 5, nous avons reporté la différence V0-V1 en distinguant les deux groupes BF et HF. La différence est calculée à partir de la proportion de chaque phonème dans son jeu de données respectif. Ainsi, dans la partie positive apparaissent les phonèmes qui sont proportionnellement plus représentés dans V0 que dans V1 (donc moins souvent réduits) et dans la partie négative les phonèmes qui sont plus présents en V1 qu'en V0. Les barres rouges représentent la présence des phonèmes en BF et les grises celles des phonèmes présents en HF.

La première observation qui émerge de la figure 5 est que la répartition des phonèmes pour la différence V0-V1 est très comparable à ce que l'on observe sur la figure 2. Cela n'est pas surprenant car il s'agit juste ici de détailler la répartition phonétique selon la fréquence des mots. On note ainsi que la différence V0-V1 est de faible ampleur en ce qui concerne les mots BF (dans tous les cas moins de 5%). Autrement dit, la réduction semble affecter globalement l'ensemble de phonèmes de la même façon (même si on note des disparités entre les deux extrémités). En revanche, les mots HF entraîne une très forte disparité concernant la proportion de phonèmes produit en V0 et en V1. Comparons les phonèmes aux deux extrémités des valeurs : /A/ n'apparaît

pas dans lœffectif de V1 HF alors quœil représente 12,6% (2ème rang) des phonèmes dans lœffectif de V0 HF. /A/ nœst donc pas réduit dans les mots HF. A lœopposé, /l/ est sur-représenté en V1 HF (21,26%) alors quœil ne représente que 6,47% des phonèmes de VO HF. Les mots HF semblent donc fonctionner comme une bascule favorisant (ou non) la réduction : les phonèmes *résistants* (ex : /A/ /s/ /k/) réduisent encore moins en HF, alors que les phonèmes *peu résistants* (ex : /l/ /y/ /w/) réduisent plus en HF. Cette observation est surprenante dans la mesure où plusieurs travaux tendent à montrer que la réduction est globalement plus importante dans les mots fréquents.

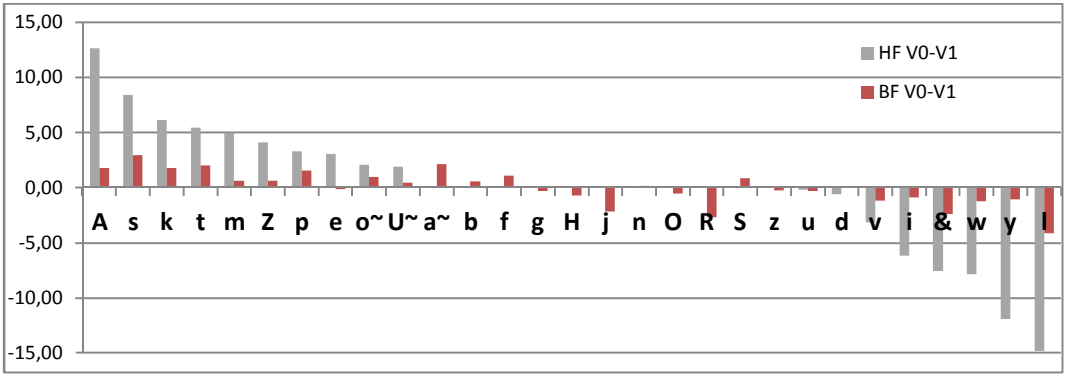


FIGURE 5: Différence (V0-V1) des fréquences dœoccurrence (en %) des phonèmes pour chaque jeu de données (V0 et V1) en fonction de la fréquence (en rouge, les BF et en gris les HF).

Quelques phonèmes montrent un effet plus important sur les mots BF que sur les mots HF. Notons particulièrement la consonne /R/ qui est très fréquente en V0 comme en V1 (3ème rang dans les deux versions). /R/ est un des très rares phonèmes à présenter un taux très important dœoccurrences réduites mais uniquement dans les mots de basse fréquence (ce qui est tout à fait cohérent avec les résultats observés en figure 4). Ce phénomène sœexplique par lœabsence de /R/ dans des mots fonctions très fréquents comme les pronoms (47 occurrences pour /R/ en V1, contre 1280 pour /l/) ou les déterminants (18 occurrences pour /R/ en V1, contre 1891 pour /l/). /R/ est donc un phonème très fréquent, mais sa fréquence nœest pas due à sa présence dans des mots fonctions fréquents, mais plutôt à sa présence très répandue dans les mots de basse fréquence. On voit ainsi que la relation entre fréquence des mots et réduction phonétique est complexe

4 Discussion

Nous avons entrepris de comparer un jeu de données comprenant lœensemble des phonèmes du corpus CID (V0) avec une sélection comprenant les phonèmes les plus courts du corpus (trame minimale : 30ms, V1). Ces deux versions étaient mises en relation avec des propriétés lexicales, syntaxiques et de fréquence. Cette comparaison nous a permis de mieux cerner à la fois la nature phonétique de la réduction, mais aussi le contexte dans lequel elle est favorisée. Nous avons vu que lœensemble des phonèmes est présent dans les deux versions. Toutefois, pour certains leur proportion augmente ou diminue en V1. Les liquides, glissantes et voyelles fermées sont proportionnellement plus présentes dans V1, tandis que les plosives, fricatives et voyelles ouvertes le sont moins. Il semble donc que lœon trouve moins de phonèmes réduits aux extrémités dœune échelle dœaperture (plosives -voyelles ouvertes), tandis quœau centre de cette échelle se trouvent les liquides et les glissantes, plus propices à la réduction. Ce résultat est cohérent avec nos connaissances sur les liquides et glissantes qui sont considérées comme des phonèmes variables et

instables (Chafcoullof, 1983). On peut donc faire l'hypothèse qu'elles sont les meilleures candidates concernant la *latitude* dont peut disposer le locuteur dans l'*optimisation* de sa production. Nous avons vu également qu'il n'y a pas de lien entre la fréquence des phonèmes et leur propension à réduire : la glissante /H/ est très peu fréquente dans le corpus, mais c'est elle qui présente le plus fort taux de réduction. Mais cette propension naturelle à réduire est également conditionnée par la nature des mots dans lesquels ces phonèmes sont produits. L'observation des phonèmes réduits au sein des CMS nous apprend ainsi que si certains phonèmes réduisent dans toutes les catégories de mots (/l/), pour d'autres la réduction est plus forte dans des CMS où ils sont le plus impliqués (verbes pour /E/, mots de contenus pour /R/). La liquide /l/ semble être la plus affectée par la réduction car elle résulte d'une combinaison de plusieurs facteurs : le /l/ est un phonème par nature court, variable et il apparait dans des catégories des mots très fréquents comme les pronoms (lui, le, la, les il, elle, ils, elles, etc.) ou les déterminants (le, la, les, etc.). Notons d'ailleurs que les pronoms « il » et « elle » sont souvent produits sous la forme réduite /i/ et /E/. Ces pronoms et déterminants sont des monosyllabes très prédictibles dans le message linguistique. La combinaison de ces deux facteurs en fait donc un excellent candidat pour la réduction. Mais l'autre liquide /R/, également très propice à la réduction, est très peu présente dans les mots de fonction fréquents. Sa forte présence en V1 est donc essentiellement due à sa présence importante dans de très nombreux mots de basse fréquence. Il semble donc que la propension d'un phonème à réduire soit due à la combinaison de facteurs multiples et diversifiés. Enfin, un phénomène surprenant est la différence entre V0 et V1 concernant les mots de haute et de basse fréquence. Les mots HF semblent fonctionner comme une bascule favorisant (ou non) la réduction. On voit que /l/ est plus présent en V1 qu'en V0, et ce phénomène est plus accentué en HF. Cela est conforme à l'idée que les mots fréquents sont plus affectés par la réduction (Jurafsky et al. 2001, Pluymaekers et al. 2005). En revanche, /A/ est moins présent en V1 qu'en V0 et ce phénomène est également plus accentué en HF, ce qui suppose que /A/ (tout comme /s/ /k/ ou /t/) réduit moins en HF qu'en BF.

Ce travail avait pour objectif d'utiliser les phonèmes extra-courts, indicateur de réduction phonétique, dans le but 1/ d'évaluer la propension de chaque phonème à réduire et 2/ de mieux cerner le contexte lexical dans lequel cette réduction s'opère. Nous avons vu que les liquides et glissantes semblent être de très bonnes candidates à la réduction, de même que la voyelle /E/, extrêmement fréquente, particulièrement lorsqu'elle est produite dans un verbe. Concernant les liquides, alors que /l/ et /R/ sont deux phonèmes très fréquents, on observe que leur propension à réduire se fait dans deux contextes très différents : les productions réduites de /l/ sont majoritairement dues à sa présence dans les déterminants et les pronoms (peu de mots répétés de très nombreuses fois), tandis que celles de /R/ sont essentiellement dues à sa présence dans les verbes et les noms (une grande quantité de mots répétés très peu de fois). On voit donc que le lien entre réduction et mots fréquents/mots fonction (Johnson, 2004) est en partie vrai, mais ce lien revêt une réalité phonétique assez complexe.

Notons enfin que ce travail porte sur l'analyse des phonèmes courts, mais pas précisément sur des *zones réduites*. Par *zones réduites* nous entendons une suite de segments de durées très courtes (Meunier, 2012). L'analyse de ces zones pourrait nous éclairer sur le fonctionnement de la réduction en lien avec des unités plus larges que le lexique (prosodie, discours) et pourrait ainsi modifier les observations faites ici sur les propriétés phonétiques de la réduction.

Remerciements

Ce travail a pu être réalisé grâce aux travaux d'enrichissement apportés au corpus du CID et réalisés dans le cadre du projet OTIM (P. Blache, ANR 2008-2011, http://www.lpl-aix.fr/~otim/).

Références

ADDA-DECKER M., GENDROT C., NGUYEN N. (2008). Contributions Du Traitement Automatique de La Parole à Lôétude Des Voyelles Orales Du Français. *Traitement Automatique Des Langues* 49 (3), 13ï 46.

BERTRAND R., BLACHE P., ESPESSER R., FERRE G., MEUNIER C., PRIEGO-VALVERDE B., AND RAUZY S. (2008). Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle. *Traitement Automatique Des Langues*, 49, 105ï 134.

BIGI B. (2015). SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech. *The Phonetician*, 111-112, 54-69.

CHAFCOULOFF M. (1983). A propos des indices de distinction /1-R/ en français. *Speech Communication*, 2(2ï 3), 137ï 139.

DUEZ D. (1995). On spontaneous French speech : aspects of the reduction and contextual assimilation of voiced stops. *Journal of Phonetics*, 23, 407-427.

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. Actes des *International Conference of Phonetic Sciences*, 1417-1420.

HALLE P., ADDA-DECKER M. (2007). Voicing assimilation in journalistic speech. Actes du *16th international congress on phonetic sciences*, 493ï 496.

JOHNSON K. (2004). Massive reduction in conversational American English. Proceedings of *the 10th International Symposium - Spontaneous Speech: Data and Analysis*. In K. Yoneyama & K. Maekawa (eds.) Tokyo, 29-54.

JURAFSKY D., BELL A., GREGORY M., RAYMOND W.D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. *Frequency and the emergence of linguistic structure*, edited by J.L. Bybee and P. Hopper (Benjamins, Amsterdam), 229-254.

LINDBLOM B. (1990). Explaining phonetic variation : a sketch of the hyper- and hypospeech theory. *Speech Production and Speech Modelling*, Hardcastle W.J., Marchal A., eds., Kluwer Academic Publishers, Dordrecht, 403-439.

MEUNIER C. (2012). Contexte et nature des réalisations phonétiques en parole conversationnelle. Actes des *Journées d'Etude sur la Parole*, Grenoble (France), 1-8.

MEUNIER C., ESPESSER R. (2011). Vowel reduction in conversational speech in French: The role of lexical factors. *Journal of Phonetics*, 39 (3), 271-278.

NEW B., PALLIER C., FERRAND L., MATOS R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE. *L'Année Psychologique*, 101, 447-462.

PLUYMAKERS M., ERNESTUS M., BAAYEN R.H. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *Journal of the Acoustical Society of America*, 118, 2561-2569.

TORREIRA F., ERNESTUS M. (2011). Vowel Elision in Casual French: The Case of Vowel /e/ in the Word Cêtait. *Journal of Phonetics* 39 (1), 50ï 58.

Réseau de neurones convolutif pour l'évaluation automatique de la prononciation

Thomas Pellegrini[1] Lionel fontan[2] Halima Sahraoui[3]
(1) IRIT - Université de Toulouse, 31062, Toulouse, France
(2) Archean Technologies, 1899 av. d'Italie, 82000, Montauban, France
(3) Octogone-Lordat - Université de Toulouse, 31058, Toulouse, France
thomas.pellegrini@irit.fr, lfontan@archean.fr, sahraoui@univ-tlse2.fr

RÉSUMÉ

Dans cet article, nous comparons deux approches d'évaluation automatique de la prononciation de locuteurs japonophones apprenant le français. La première, l'algorithme standard appelé *Goodness Of Pronunciation* (GOP), compare les vraisemblances obtenues lors d'un alignement forcé et lors d'une reconnaissance de phones sans contrainte. La deuxième, nécessitant également un alignement préalable, fait appel à un réseau de neurones convolutif (CNN) comme classifieur binaire, avec comme entrée des trames de coefficients spectraux. Les deux approches sont évaluées sur deux phonèmes cibles /R/ et /v/ du français, particulièrement difficiles à prononcer pour des Japonophones. Les paramètres du GOP (seuils) et du CNN sont estimés sur un corpus de parole lue par des locuteurs natifs du français, dans lequel des erreurs de prononciation artificielles sont introduites. Un gain de performance relatif de 13,4% a été obtenu avec le CNN, avec une précision globale de 72,6%, sur un corpus d'évaluation enregistré par 23 locuteurs japonophones.

ABSTRACT

CNN-based automatic pronunciation assessment of Japanese speakers learning French

In this paper, we compare two approaches for the automatic evaluation of the pronunciation of Japanese speakers speakers learning French. The first one, the standard algorithm called Goodness Of Pronunciation (GOP), compares likelihoods obtained during forced alignment and during phone recognition with no constraint. The second, also requiring a signal-to-phone alignment, uses a convolutional neural network (CNN) as a binary classifier, with frames of spectral coefficients as input. Both approaches are evaluated on two target French phonemes /R/ and /v/, particularly difficult to pronounce for Japanese-speaking natives. GOP decision thresholds and CNN parameters are estimated on a read speech corpus of French native speakers, in which artificial pronunciation errors are introduced. A 13.4% performance gain relative was obtained with the CNN, with an overall accuracy of 72.6%, on an evaluation corpus recorded by 23 Japanese native speakers.

MOTS-CLÉS : évaluation automatique de la prononciation, réseau de neurones convolutif, français langue étrangère.

KEYWORDS: Automatic pronunciation assessment, convolutional neural network, CNN, Goodness-of-Pronunciation, French as a second language.

1 Introduction

Les systèmes d'apprentissage d'une langue seconde assisté par ordinateur tentent d'évaluer automatiquement la prononciation pour aider les apprenants. Pour l'évaluation au niveau segmental, une approche standard consiste à attribuer un score de prononciation à chaque réalisation de phone (Eskenazi, 2009). Deux grands types d'approches peuvent être distingués : 1) les approches fondées sur des scores de systèmes de reconnaissance, bruts (Sevenster *et al.*, 1998), ou relatifs, sous forme de rapports de vraisemblances comme c'est le cas de l'algorithme *Goodness Of Pronunciation* (GOP) (Witt, 1999; Kanters *et al.*, 2009) utilisé dans la présente étude, 2) les approches « signal », qui font appel à des classifieurs prenant en entrée des paramètres acoustiques (Strik *et al.*, 2007).

Dans (Strik *et al.*, 2007) par exemple, les auteurs comparent ces deux types d'approches pour distinguer les deux phonèmes /k/ et /x/ sources de confusion pour les apprenants du néerlandais. Ils obtiennent des performances légèrement meilleures en utilisant une analyse discriminante linéaire (LDA) avec des paramètres acoustiques. Cette approche présente néanmoins le désavantage d'être spécifique à chaque phone que l'on souhaite évaluer et il faut ré-estimer les poids de la LDA à chaque nouveau phone cible. Dans l'étude présentée dans cet article, nous comparons également les deux types d'approches, en utilisant un réseau de neurones convolutif (CNN), qui, outre ses performances remarquables en reconnaissance automatique de la parole, permet d'évaluer plusieurs phones simultanément.

Plus précisément, nous comparons les approches suivantes : 1) une variante plus robuste de l'algorithme de base GOP, appelée f-GOP, proposée par Luo *et al.* (2010), avec et sans seuils (utilisation d'une régression logistique), 2) un réseau de neurones convolutif qui prend en entrée des paramètres acoustiques (coefficients F-BANK) ainsi que l'identité du phone attendu, et qui prend une décision binaire trame à trame d'acceptation ou de rejet d'une prononciation. Une décision finale pour un phone est prise par un simple vote majoritaire sur l'ensemble des trames du phone.

Les expériences d'évaluation ont été réalisées sur le corpus PHON-IM, corpus de parole produite dans une tâche de répétitions de mots, recueillie auprès de 23 locuteurs japonais qui apprennent le français comme langue étrangère (FLE). Nous avons centré cette étude sur deux phones particulièrement difficiles à maîtriser pour les japonophones : [R] et [v], souvent confondus avec [l] et [b] respectivement (Tomimoto & Takaoka, 2008; Yamasaki & Hallé, 1999). La taille de ce corpus étant très petite pour servir à la fois à l'apprentissage et à l'évaluation des méthodes, nous avons utilisé le corpus de parole lue BREF80, enregistré par 80 locuteurs français natifs. Des erreurs de prononciation sont simulées en introduisant des substitutions de phones dans le dictionnaire de prononciation utilisé pour réaliser les alignements. Si cette méthode a été appliquée avec succès dans Kanters *et al.* (2009), ce travail nous a permis d'en déceler des limites que nous décrirons.

2 Approches

Les approches utilisées nécessitent d'aligner le signal de parole avec les séquences de phones attendus. Nous avons utilisé pour cela des modèles acoustiques de phones indépendants du contexte (39 monophones), plus adaptés que des modèles dépendants du contexte pour des applications de détection d'erreurs de prononciation (Kawahara & Minematsu, 2012). Ces modèles sont des HMM gauche-droite à trois états avec des mélanges de Gaussiennes à 32 composantes, entraînés sur le corpus ESTER phase I (de Calmès *et al.*, 2005). Ce sous-corpus contient 31 heures de parole de diverses émissions de radio française. Les modèles sont disponibles en ligne (Farinas, 2013).

2.1 Algorithme *f-GOP*

Cette méthode a été proposée au départ pour l'évaluation de prononciation non-native (Witt & Young, 2000; Kanters *et al.*, 2009; Luo *et al.*, 2010), et a été également utilisée avec succès pour caractériser les troubles pathologiques de production de la parole dans des cas de sévérité modérée (Pellegrini *et al.*, 2015).

L'algorithme GOP de référence peut être décomposé en trois étapes : 1) la phase d'alignement forcé d'une séquence de phones attendus (parole lue) au signal de parole, 2) la phase de reconnaissance de phones sans contrainte et 3) le calcul des scores comme la différence entre les log-vraisemblances des deux phases précédentes pour chaque phone aligné. Les scores varient entre 0 et 10 environ, et plus la valeur est grande, plus une erreur de prononciation est susceptible d'avoir été déctectée. Les ordres de grandeur des vraisemblances dépendent entre autres du phone considéré. Pour cette raison, il est commun de déterminer à partir d'un sous-corpus de développement les seuils de décision pour chaque phone cible.

Dans ce travail, nous utilisons une version plus performante du GOP, appelée f-GOP, qui force la phase de reconnaissance libre à utiliser les segments trouvés lors de l'alignement (Luo *et al.*, 2010). Les seuils d'acceptabilité des phones [R] et [v] ont été estimés à 1,13 et 2,97 respectivement. Pour déterminer ces seuils, nous avons utilisé BREF80, le corpus de parole lue par des locuteurs français natifs, que nous décrivons à la section 3. Les locuteurs de ce corpus étant des francophones natifs, nous considérons toutes les occurrences des deux phones comme acceptables (classe positive). Pour simuler des erreurs de prononciation, nous considérons que toutes les occurrences de [l] et [b] correspondent à des prononciations erronées de /R/ et /v/ respectivement. Pour ce faire, nous forçons le système d'alignement à utiliser un [R] à la place des [l] en remplaçant le phone [l] par [R] dans le lexique de prononciation (même chose pour [b] et [v]).

2.2 Régression logistique (f-GOP+RL)

Pour éviter de devoir fixer des seuils de décision, nous avons utilisé un classifieur fondé sur une régression logistique (RL). Très populaire en traitement du langage naturel, cette technique obtient des performances similaires aux séparateurs à vaste marge (Theodoridis, 2015), avec l'avantage de mettre en jeu des poids θ qui ont une interprétation sur l'importance des paramètres d'entrée. Pour pouvoir comparer les performances du modèle RL avec f-GOP, nous avons utilisé comme paramètres d'entrée uniquement les scores f-GOP et l'identité du phone attendu. Les poids du modèle sont entraînés sur les mêmes exemples de BREF80 qui ont servi à fixer les seuils des scores f-GOP. On constate que le poids du score GOP après apprentissage vaut -0,633, une valeur négative qui correspond bien au fait que plus un score GOP est élevé, plus une erreur de prononciation est vraisemblable. Les poids attribués aux paramètres catégoriels d'identité du phone sont 0,627 et 0,445 pour /v/ et /R/ respectivement. Le poids pour /v/ est légèrement plus grand que celui de /R/, ce qui est également cohérent avec le fait que le seuil GOP trouvé précédemment est plus élevé pour ce phone.

2.3 Réseau de neurones convolutif (CNN)

La figure 1 illustre l'architecture du réseau mis en place pour cette étude. Il comporte une couche d'entrée composée d'une trame de 40 coefficients log-F-BANK calculés sur une fenêtre de 20ms, à

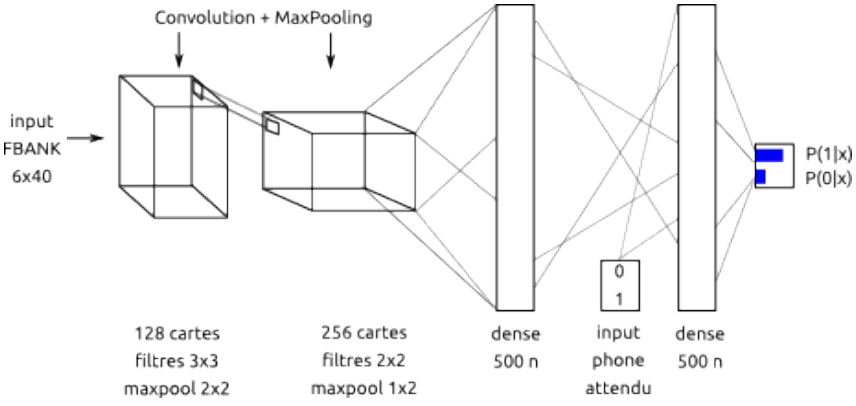


FIGURE 1 – Architecture du réseau convolutif.

laquelle ont été ajoutées les trois trames voisines précédentes et les deux trames voisines suivantes pour un total de six trames. Deux trames voisines sont séparées par 10ms. Deux couches de convolution avec sous-échantillonnage (*Max pooling*) permettent d'obtenir respectivement 128 et 256 cartes d'activation qui servent de paramètres d'entrée à deux couches cachées denses de 500 neurones avec une fonction d'activation ReLu. L'information de l'identité du phone attendu, connue grâce à un alignement forcé préalable (GOP, phase 1), est donnée à la dernière couche cachée de décision, sous la forme d'un vecteur de zéros avec un unique 1 (représentation *one-hot*). Cette information est nécessaire puisqu'un seul modèle est utilisé pour tous les phones d'intérêt (ici au nombre de deux : [R] et [v]). Une décision par trame est prise et la décision globale pour un phone est prise à l'aide d'un vote majoritaire sur l'ensemble des trames qui le compose. Les poids du modèle ont été initialisés à l'aide de la méthode « Xavier » (Glorot & Bengio, 2010), et entraînés avec une descente de gradient momentum Nesterov, avec une fonction de coût de type entropie binaire croisée. La méthode de régularisation *dropout* ($p = 0,5$) n'est utilisée qu'avec les couches cachées denses. D'autres architectures ont été testées, avec une seule couche de convolution ou avec trois couches de convolution par exemple, et celle présentée ici a donné les meilleurs résultats. Pour la mise en œuvre de ces modèles, nous avons utilisé les boîtes à outils Theano (Bergstra *et al.*, 2010) et Lasagne [1].

Pour entraîner le modèle, nous avons divisé le corpus BREF80, décrit dans la section suivante, en deux sous-corpus *Train* et *Val* dans les proportions 90% / 10%, soit 300K / 30K exemples, respectivement. La figure 2 montre l'évolution du coût sur *Train* et *Val*, ainsi que la précision obtenue sur *Val*, au cours des 100 premières itérations d'apprentissage. Une droite horizontale situe la performance obtenue sur PHON-IM par le modèle final : 88,9% de classification correcte des trames. Si le coût sur *Train* continue à diminuer après 100 itérations, la performance sur *Val* atteint un plateau rapidement. Le critère d'arrêt que nous avons utilisé était une diminution minimale de 1e-3 sur le coût calculé sur *Val* au cours de trois itérations successives, ce qui a conduit à un nombre de 178 itérations.

1. https://github.com/Lasagne/Lasagne

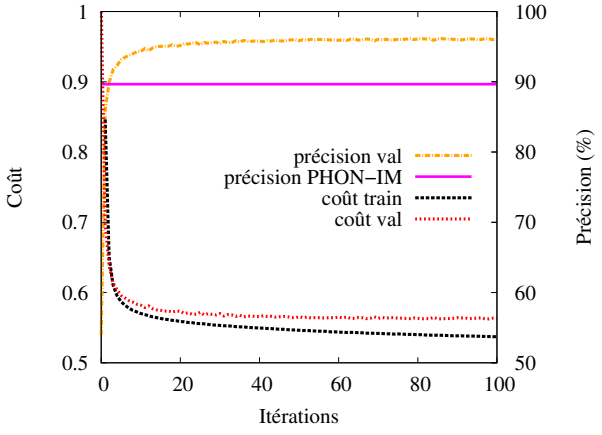


FIGURE 2 – Évolution du coût sur *train*, du coût et précision sur *val*, au cours des 100 premières itérations d'apprentissage. La performance sur le corpus de test PHON-IM est indiquée par une droite horizontale (88,9%).

corpus	BREF80		PHON-IM	
	correct	incorrect	correct	incorrect
/R/	21K	16K (phone [l])	215	128
/v/	5K	3K (phone [b])	267	50

TABLE 1 – Nombres d'occurrences de /R/ et de /v/ dans BREF80 et PHON-IM.

3 Corpora

3.1 BREF80

Le corpus BREF est un corpus de plus de 100 heures de parole lue, recueillie auprès de 120 locuteurs français natifs, qui ont lu des textes du journal *Le Monde* (Gauvain *et al.*, 1990). Nous en avons utilisé une sous-partie, appelée BREF80, qui correspond aux enregistrements de 80 locuteurs.

Le nombre d'occurrences des deux phones cibles [R] et [v] dans ce corpus est indiqué dans le tableau 1 : 21K et 5K. Toutes ces occurrences sont considérées comme des réalisations correctes. Les nombres d'occurrences incorrectes correspondent aux nombres de [l] et [b] du corpus, 16K et 3K, considérés comme des exemples de la classe négative. Comme dit dans la section précédente, ce corpus a été réparti en deux sous-corpus *Train* et *Val* dans les proportions 90% / 10%, spécifiquement pour l'apprentissage du modèle CNN. Cette division est nécessaire pour établir un critère d'arrêt sur les itérations de descente de gradient.

TABLE 2 – Corpus PHON-IM : jugements d'acceptabilité par phonème et par position

	/v/		/R/	
	Acceptable	Non accept.	Acceptable	Non accept.
Initiale	89	30 (25,2%)	53	59 (52,7%)
Intervocalique	75	6 (7,4%)	56	55 (49,6%)
Finale	103	14 (12,0%)	106	14 (11,7%)
Total	277	50 (15,8%)	215	128 (37,3%)

3.2 PHON-IM

3.2.1 Enregistrements utilisés dans le cadre de cette étude

PHON-IM est un projet de recherche visant à étudier les compétences de perception et de production phonétique chez des apprenants japonophones de FLE, et ce dans une perspective longitudinale. Il s'inscrit dans le cadre d'un programme annuel d'échange d'étudiants entre l'Université Ritsumeikan de Kyoto et le Département de FLE (DEFLE) de l'Université Toulouse II – Jean Jaurès. Chaque année un groupe d'apprenants débutants (A1/A2) vient à Toulouse pour un séjour d'immersion linguistique d'un mois. Les réalisations cibles /R/, /l/, /v/ et /b/ présentent typiquement des difficultés particulières pour les locuteurs natifs japonophones en écoute et prononciation du français langue étrangère.

Pour ce travail nous avons utilisé un sous-ensemble d'enregistrements collectés auprès de 23 apprenants japonais, dont la tâche était de répéter des mots ou pseudo-mots dissyllabiques contenant les phonèmes /R/ et /v/. Les deux phonèmes apparaissent à la fois dans les positions initiale, intervocalique et finale, la fréquence d'occurrence dans ces trois positions étant équilibrée au sein du corpus. Au total, le sous-ensemble comprend 368 réalisations de /v/ et 414 réalisations de /R/.

3.2.2 Annotations par des enseignants de FLE

Deux enseignants de FLE ont évalué les 782 réalisations, en indiquant si selon eux la réalisation était acceptable en fonction de la cible attendue. L'accord entre les deux annotateurs est de 84,4% ; cet accord est légèrement supérieur pour les réalisations du phonème /R/ (86,2%) que pour celles du phonème /v/ (82,9%).

En cas de réalisation non acceptable, les deux annotateurs étaient enjoints à indiquer quel était le phone se rapprochant le plus de la réalisation de l'apprenant. Pour le phonème /R/, les réalisations jugées non acceptables ont le plus souvent été décrites comme proches du phone [h] présent dans le système phonético-phonologique du japonais. Pour les réalisations du phonème /v/, c'est la fricative bilabiale [β] qui a le plus souvent été évoquée par les deux annotateurs, et dans une moindre mesure l'occlusive [b].

Au total, parmi les 660 occurrences pour lesquelles un accord inter-annotateur a été obtenu, 15,8% des réalisations de /v/ ont été jugées non acceptables, contre 37,3% pour le /R/. La table 2 décrit les

		Correctement Acceptés			Correctement Rejetés		
Modèles	tx global	prec.	rappel	F1	prec.	rappel	F1
f-GOP	68,5/58,7	73,2/91,5	78,6/56,2	75,8/69,6	58,9/23,5	51,6/72,0	55,0/35,4
f-GOP+RL	71,1/57,1	71,6/92,3	89,3/53,6	79,5/67,8	69,3/23,5	40,6/76,0	51,2/35,9
CNN	68,5/77,0	71,1/91,1	83,7/80,5	76,9/85,5	61,1/35,8	43,0/58,0	50,5/44,3

TABLE 3 – Résultats obtenus sur le corpus de test PHON-IM. Dans chaque cellule, les pourcentages sont donnés pour les phonèmes /R/ et /v/, respectivement.

scores par position. Il semble que ce soit la position initiale qui pose le plus de difficultés, de manière beaucoup plus marquée pour /v/ que pour /R/, avec 30 occurrences jugées non-acceptables pour cette position, contre 6 et 14 en position intervocalique et finale respectivement.

4 Résultats

Dans le tableau 3 donne les mesures de performance obtenues sur le corpus d'évaluation PHON-IM, avec le détail pour les deux phonèmes /R/ et /v/ dans chaque cellule du tableau. De manière globale, les performances f-GOP et f-GOP+RL sont très proches, ce qui est confirmé par le fait que leurs prédictions sont identiques à 89,2% pour /R/ et à 97,2% pour /v/. Ce résultat était attendu dans la mesure où les mêmes informations sont utilisées en entrée, à savoir le score GOP et l'identité du phone attendu. Le modèle de régression linéaire apprend de manière autonome les seuils d'acceptabilité, ce qui confirme l'intérêt d'utiliser un tel classifieur.

En revanche, les prédictions obtenues avec le CNN différent de manière significative : les pourcentages de prédictions identiques avec le modèle f-GOP+LR tombent à 76,4% et 50,5% des cas pour /R/ et /v/, respectivement. La première colonne du tableau donne le taux global de bonne classification, i.e. le ratio du nombre de réalisations correctement acceptées (CA) ou rejetées (CR) sur le nombre d'occurrences total. Les trois approches ont donné un taux similaire pour /R/, autour de 70%. En revanche, le CNN a un taux bien meilleur pour /v/ : 77,0% contre 58,7% et 57,1% pour f-GOP et f-GOP+RL. Cela est essentiellement dû à un rappel des CA meilleur avec CNN, de 80,5%, par rapport aux rappels de f-GOP, 56,2%, et de f-GOP+RL, 53,6%. f-GOP et f-GOP+RL ont tendance à rejeter des occurrences de /v/ qui ont été jugées correctes par les annotateurs. Cette tendance est également visible de par les taux de rappel des occurrences CR, qui sont meilleurs pour ces deux approches, mais qui ne peuvent compenser les valeurs plus faibles des rappels CA et de la précision CR (23,5%).

Pour analyser ces résultats plus en détails, la figure 3 illustre les prédictions faites par le CNN en fonction de la position du phone dans le mot (intiale, intervocalique ou finale) pour /v/ (à gauche) et pour /R/ (à droite). Deux histogrammes de trois barres chacun sont donnés pour chaque position : les occurrences « acceptées » et les occurrences « rejetées ». Les trois barres correspondent de gauche à droite : à l'annotation manuelle, aux nombres d'occurrences correctement (CA et CR) et incorrectement (FA et FR) classées par le CNN. Il est intéressant de noter que pour les deux phones, c'est la position initiale qui est la source du plus grand nombre d'erreurs : les rejets incorrects pour /v/, les acceptations incorrectes pour /R/. La position intervocalique est celle qui a les meilleures performances. Ces différences de qualité des prédictions indiquent que la méthode de simulation d'erreurs devrait prendre en compte la position des phones. Ces résultats semblent montrer que les difficultés de prononciation de /v/ et /R/ des Japonophones ne sont pas les mêmes selon la position du

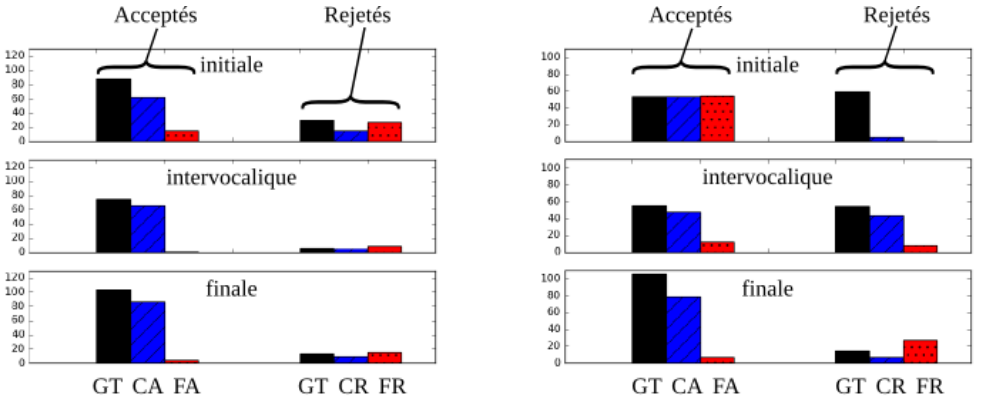


FIGURE 3 – Nombre d'occurrences de /v/ (à gauche) et de /R/ (à droite) acceptées ou rejetées par les annotateurs (noir, GT pour *ground-truth*), correctement acceptés ou rejetés par le CNN (bleu hachuré, CA et CR), et faussement acceptés ou rejetés par le CNN (rouge avec points, FA et FR).

phone, ce qui transparaît dans le fait que les annotateurs ont rejetés moins d'occurrences des phones en position intervocalique que dans les deux autres positions cumulées.

5 Conclusions

Nous avons comparé deux approches d'évaluation automatique de la prononciation de locuteurs japonophones apprenant le français : l'algorithme f-GOP et un réseau de neurones convolutif. Elles ont été évaluées sur deux phones cibles [R] et [v] du français, particulièrement difficiles à prononcer pour des locuteurs japonophones débutants. Un gain de performance relatif de 13,4% a été obtenu avec le CNN, avec une précision globale de 72,6%, sur un corpus de mots recueilli auprès de 23 locuteurs japonophones.

Pour pouvoir mettre en place ces méthodes, nous avons dû recourir à la simulation d'erreurs de prononciation dans un corpus de parole native pour pallier le manque de données de parole d'apprenants. Nous avons constaté que cela présente des limites dans la mesure où des connaissances *a priori* sont nécessaires sur les confusions les plus fréquentes faites par les apprenants. De plus, les différences de performance en fonction de la position intermot des consonnes montrent que simuler les erreurs sans en tenir compte est une approximation qui peut être améliorée.

Nous envisageons de comparer les résultats présentés ici avec deux autres situations : 1) avec une prise en compte plus fine des confusions faites en fonction de la position des phones pour la simulation des erreurs, 2) à l'inverse, sans connaissance *a priori* sur les confusions fréquentes faites par les locuteurs. Dans la deuxième situation, nous envisageons d'utiliser des occurrences de phones du français choisis aléatoirement pour simuler des erreurs, en limitant le nombre d'occurrences pour obtenir un jeu d'apprentissage équilibré. Des améliorations du modèle CNN lui-même sont également envisagées. Enfin, une nouvelle collection de données d'apprenants japonophones est actuellement en cours, dans les mêmes conditions que celles du corpus PHON-IM, ce qui permettra de doubler la taille du corpus d'évaluation.

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 expression compiler. In *Proc. of the Python for Scientific Computing Conference (SciPy)*.

DE CALMÈS M., FARINAS J., FERRANÉ I. & PINQUIER J. (2005). Campagne ESTER : une première version d'un système complet de transcription automatique de la parole grand vocabulaire (Atelier ESTER, Avignon, 30/03/2005-31/03/2005). `http://www.afcp-parole.org/camp_eval_systemes_transcription/private/atelier_mars_2005/pdf/IRIT_ester2005.pdf`. [Online ; accessed 10-April-2016].

ESKENAZI M. (2009). An overview of spoken language technology for education. *Speech Communication*, **51**(10), 832–844.

FARINAS J. (2013). Multilingual phonetic decoders. `http://www.irit.fr/recherches/SAMOVA/pagedap.html`. [Online ; accessed 20-September-2015].

GAUVAIN J.-L., LAMEL L. & ESKENAZI M. (1990). Design considerations and text selection for BREF, a large french read-speech corpus. In *Proc. ICSLP-90*, p. 1097–2000.

GLOROT X. & BENGIO Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'10)*.

KANTERS S., CUCCHIARINI C. & STRIK H. (2009). The Goodness of Pronunciation Algorithm : A Detailed Performance Study. In *SLaTE 2009 - 2009 ISCA Workshop on Speech and Language Technology in Education*, p. 2–5.

KAWAHARA T. & MINEMATSU N. (2012). *Tutorial on CALL Systems at Interspeech*. Portland.

LUO D., QIAO Y., MINEMATSU N., YAMAUCHI Y. & HIROSE K. (2010). Regularized-MLLR speaker adaptation for computer-assisted language learning system. In *Proc. Interspeech*, p. 594–597, Makuhari.

PELLEGRINI T., FONTAN L., MAUCLAIR J., FARINAS J., ALAZARD-GUIU C., ROBERT M. & GATIGNOL P. (2015). Automatic Assessment of Speech Capability Loss in Disordered Speech. *ACM Trans. Access. Comput.*, **6**(3), 8 :1–8 :14.

SEVENSTER B., KROM G. D. & BLOOTHOOFT G. (1998). Evaluation and training of second-language learners' pronunciation using phoneme-based HMMs. In *Proc. STiLL*, p. 91–94, Marholmen.

STRIK H., TRUONG K. P., DE WET F. & CUCCHIARINI C. (2007). Comparing classifiers for pronunciation error detection. In *Proc. INTERSPEECH*, p. 1837–1840.

THEODORIDIS S. (2015). *Machine Learning*. Elsevier.

TOMIMOTO J. & TAKAOKA Y. (2008). Le français, une langue imprononçcable pour les Japonais ? *Rencontres Pédagogiques du Kansaï*.

WITT S. (1999). *Use of Speech Recognition in Computer-Assisted Language Learning*. Phd dissertation, University of Cambridge, Dept. of Engineering.

WITT S. & YOUNG S. (2000). Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning. *Speech Communication*, **30**, 95–108.

YAMASAKI H. & HALLÉ P. (1999). How do native speakers of japanese discriminate and categorize french /r/ and /l/ ? In *Proceedings of ICPhS*, p. 909–912, San Francisco.

Rôle des contextes lexical et post-lexical dans la réalisation du schwa : apports du traitement automatique de grands corpus

Yaru WU[1] Martine ADDA-DECKER[1, 2] Cécile FOUGERON[1]
(1) UMR 7018 - Laboratoire de phonétique et phonologie (LPP), 19 rue des Bernardins, Paris, France
(2) Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur (LIMSI), Rue John Von Neumann, Orsay, France
{yaru.wu,martine.adda-decker,cecile.fougeron}@univ-paris3.fr

RÉSUMÉ

Le rôle du contexte est connu dans la réalisation ou non du schwa en français. Deux grands corpus oraux de parole journalistique (ETAPE) et de parole familière (NCCFr), dans lesquels la realisation de schwa est déterminée à partir d'un alignement automatique, ont été utilisés pour examiner la contribution du contexte au sein du mot contenant schwa (lexical) vs. au travers de la frontière avec le mot précédent (post-lexical). Nos résultats montrent l'importance du contexte pré-frontière dans l'explication de la chute du schwa dans la première syllabe d'un mot polysyllabique en parole spontanée. Si le mot précédant se termine par une consonne, nous pouvons faire appel à la loi des trois consonnes et au principe de sonorité pour expliquer des différences de comportement en fonction de la nature des consonnes en contact.

ABSTRACT

Role of lexical and post-lexical contexts in French schwa realisations : benefits of automatic processing of large corpora

The role of context is known to affect the realization/deletion of schwa in French. Two large corpora of public journalistic speech (ETAPE) and casual speech (NCCFr), in which schwa realization is defined via automatic forced alignment, are used to examin the contribution of context both within word (lexical) and across word boundaries (post-lexical). Our results highlight the role of the pre-boundary context in the observed schwa deletion rates. If the preceding word ends with a consonant, predictions follow the 3 consonants rule and sonority principles to explain the differences observed in schwa deletion rates.

MOTS-CLÉS : chute de schwa, grands corpus, parole spontanée, contexte pré-frontière, alignement forcé .

KEYWORDS: schwa deletion, large corpora, spontaneous speech, pre-boundary context, forced alignment.

1 Introduction

Depuis Grammont (Grammont, 1894), le schwa est décrit comme une voyelle instable dont la réalisation ou non-réalisation est fonction, entre autres, du contexte consonantique environnant. La présence du schwa deviendrait obligatoire dès lors que le cluster consonantique résultant de son élision potentielle inclut au moins trois consonnes. Grammont est ainsi à l'origine de la loi des

3 consonnes (L3C), qui prédit la présence obligatoire du schwa dans ce cas. Elle a été reprise et étudié par de nombreux phonéticiens et phonologues depuis (Durand *et al.*, 2009). En dépit d'une loi simple, la situation en production n'est cependant pas si simple. Grammont a lui-même décrit des exceptions en proposant par exemple que les clusters formés d'obstruante et liquide (par exemple [pl]) ne comptent que pour une seule consonne par rapport à cette loi (voir par exemple (Delattre, 1944, 1966; Lyche, 1993)). De plus, on sait qu'en plus des limitations pouvant être liées au nombre de consonnes successives, les restrictions dans les suites consonantiques permises dans les langues (et donc les propriétés phonotactiques de celles-ci), sont fonction de la nature des consonnes adjacentes, notamment en terme de rapport de sonorité (Clements, 1990). Ainsi la formation de groupe de consonne en attaque de syllabe privilégiera une suite à sonorité croissante jusqu'au noyau. Dans cette étude, nous nous intéresserons au poids de ces différentes contraintes (nombre et nature/sonorité des consonnes) sur la chute potentielle de schwa dans des cas où les consonnes pouvant être mises en contact lors de la chute de schwa appartiennent soit au même mot soit à des mots différents. Pour cela, nous étudions des mots polysyllabiques contenant un schwa en syllabe initiale d'un mot, #C_1_$C_2(C_3)V$ position dans laquelle la délétion de schwa est fréquente (Côté, 2000), en fonction du nombre et de la nature des consonnes en contexte aussi bien au sein du mot (ex. s(e)crétaire = 3C [skR], f(e)nêtre =2C [fn]), qu'au travers de la frontière de mot en considérant comment le mot précédent se termine (ex. la s(e)crétaire = 3C [a#skR] vs. bonne s(e)crétaire = 4C [n#skR]). L'accès à de grands corpus transcrits et alignés (en segments phonétiques) nous permet de localiser les contextes consonantiques susceptibles de générer ou d'élider un schwa à l'intérieur des mots et à leur frontière.

Dans la première partie de notre étude, nous étudierons les contraintes phonotactiques lexicales (interne au mot). Nous limitons notre étude à deux types de formes :

- Forme$_{2C}$: #C_1_C_2V
- Forme$_{3C}$: #C_1_C_2C_3V

avec "#" frontière de mot ; "_" position de schwa ; V : voyelle pleine. Nous ne considérons pas la séquence #C_1C_2_C_3V trop peu fréquente dans nos données (ex. prenons). Ne rentrent pas non plus dans notre étude tous les mots monosyllabiques avec schwa comme noyau (le, de, ce, se, ...) extrêmement fréquents dans les corpus. Dans la deuxième partie, nous examinerons ces mêmes contraintes phonotactiques mais sur un domaine d'application post-lexical, i.e. au travers des frontières de mots. Nous comparerons deux types de contexte pré-frontière, c'est à dire précédant le mot contenant le schwa : le mot précédant se termine par une voyelle pleine (V#) ou par une consonne (C#). Les cas où le mot avec schwa est précédé d'une pause ou hésitation (cas relativement peu fréquent dans nos données) ne seront pas examinés en détail.

2 Méthodologie

Les approches et méthodes développées pour la reconnaissance automatique de la parole ont donné, entre autres, les systèmes d'alignement forcé, qui consistent à localiser les mots, ainsi que les segments qui les composent dans le flux de parole continu, à condition de disposer d'une transcription orthographique préalable. Les alignements correspondants peuvent servir pour élaborer ou pour valider des hypothèses linguistiques en production.

Afin de quantifier automatiquement les variantes majeures dans la parole, on peut faire appel à l'alignement automatique. En particulier pour les variantes concernant le schwa (avec ou sans schwa réalisé), il faut expliciter celles-ci dans le dictionnaire de prononciation. Si l'alignement automatique

peut avoir des erreurs ne permettant pas de localiser et identifier toutes les variantes effectivement produites, il permet néanmoins de faire émerger des tendances et surtout permet de traiter une grande quantité de parole (Bürki *et al.*, 2008). Cette approche est donc différente de celle suivie par le projet PFC (Durand *et al.*, 2003), par exemple, dans lequel la présence de schwa est déterminée à l'écoute.

Les données ont été alignées automatiquement par le système de reconnaissance automatique du LIMSI (Gauvain *et al.*, 2005). Afin de déterminer quels mots contiennent un schwa sous-jacent, nous utilisons la transcription phonologique de Lexique380 (New *et al.*, 2007) comme référence. Nous limitons notre étude à l'intersection des mots à schwa trouvés à la fois dans Lexique380 et dans nos corpus. Par exemple, les noms propres fréquents dans les corpus ne sont pas inclus dans Lexique.

Nous utilisons deux grands corpus dans notre étude incluant environ 80 heures de données de parole continue : (1) le corpus du projet ETAPE (Gravier *et al.*, 2012) qui contient 13 heures et demie de données radiophoniques et 29 heures de données télévisuelles incluant des débats et des conversations assez libres auquel ont été ajouté environ 15 heures du corpus radiophonique ESTER (Galliano *et al.*, 2005) ; (2) le corpus NCCFr (Torreira *et al.*, 2010), qui contient 35 heures de conversations familières entre amis.

Ces données ont été segmentées automatiquement en mots et en phonèmes par le système d'alignement forcé du LIMSI. Nous pouvons qualifier la transcription segmentale résultante de transcription phonétique dans la mesure où le système prévoit des prononciations avec variantes (avec et sans schwa) : cela veut dire que si un mot a été réalisé sans schwa entre deux consonnes, l'alignement forcé peut aligner directement les deux consonnes comme segments consécutifs sans forcer un segment schwa entre les deux - ce qui serait le cas sans variante. Nous pouvons donc supposer que la transcription segmentale résultante est proche de la prononciation des locuteurs. Il serait intéressant de tester si ces résultats restent stables ou s'ils varient, même légèrement, avec un autre système.

La transcription donnée par Lexique (New *et al.*, 2007), par opposition, nous donne la forme phonologique de référence puisqu'elle note si le mot contient un schwa. C'est donc en regardant la différence de transcription au niveau de la présence du schwa entre Lexique (forme théorique sous-jacente) et les formes alignées de ETAPE/NCCFr (productions des locuteurs) que nous déterminons si le schwa a été produit ou non. Nous avons tout d'abord généré une liste de prononciation de référence pour chacun des mots présents à la fois dans Lexique et dans le corpus ETAPE ou NCCFr. Ainsi, chaque mot retenu est accompagné de sa prononciation provenant de Lexique (forme phonologique) et de sa prononciation alignée (production du locuteur) permettant de catégoriser l'occurrence comme ayant un schwa 'présent' ou 'absent'(catégories que nous utiliserons dans les figures suivantes) et donc un taux de réalisation/chute de schwa.

La table 1 résume les conditions testées et la terminologie adoptée. Nous mesurons d'abord les taux de réalisation/chute de schwa à l'intérieur des mots tous contextes pré-frontière confondus. Nous ne considérons que les schwas en syllabe initiale de mots polysyllabiques et ceci dans deux types de formes, Forme$_{2C}$ et Forme$_{3C}$: la position de schwa est précédée de 1 consonne et suivie de 1 ou 2 consonnes. Nous éclatons ensuite ces contextes génériques suivant la nature des consonnes afin d'étudier le lien entre les taux de chute de schwa et la sonorité croissante ou non de la séquence qui résulterait de la chute du schwa. Afin de déterminer la sonorité des catégories, nous retenons l'échelle de sonorité simplifiée suivante :

$$Sn(C') \leq Sn(L) \leq Sn(G) \leq Sn(V)$$

où C' représente les non-approximantes (plosives, nasales, fricatives), L les liquides, G les glides et V les voyelles. Partant de l'hypothèse que lors de la chute de schwa, les consonnes en présence se

resyllabent avec la voyelle suivante, nous regarderons si la chute de schwa est préférée lorsque les consonnes forment une séquence de sonorité croissante (i.e. respectant le principe de sonorité, ☺) plutôt que décroissante (☹) ou en plateau (☹). Afin de comparer les contextes sur un nombre suffisant de données, nous ne retenons que les clusters avec au moins 500 occurrences (voir partie 3.1) dans les deux corpus. Pour cette raison, seuls les séquences contenant des C' et L dans les formes Forme$_{2C}$ présentées dans le tableau ont été examinées.

Ensuite, nous étendons nos analyses au niveau post-lexical en examinant les taux de réalisation du schwa suivant que le mot précédent se termine avec une voyelle pleine (V#) ou une consonne (C#).

Condition		Nombre de consonnes considérées	Séquences en fonction de l'échelle de sonorité	
Contraintes phonotactiques lexicales			$S(C_1) < S(C_2)$?	
Intra-mot	Forme$_{2C}$:	/ə/ | **#C$_1$__C$_2$V**	#C'__C'V ☺ #C'__LV ☺ #L__C'V ☺	petite, semaine sereine, pelouse revue, leçon
	Forme$_{3C}$:	/ə/ | **#C$_1$__C$_2$C$_3$V**	**NA (pas assez de tokens)**	chevron, secrète
Contraintes phonotactiques post-lexicales				
Inter-mot		Contexte pré-frontière = V		
	Forme$_{2C}$:	/ə/ | **V#C$_1$__C$_2$V**	V#C'__C'V ☺ V#C'__LV ☺	les petites, la semaine la sereine, la pelouse
	Forme$_{3C}$:	/ə/ | **V#C$_1$__C$_2$C$_3$V**	**NA (pas assez de tokens)**	la revue, des leçons
		Contexte pré-frontière = C		
	Forme$_{2C}$:	/ə/ | **C#C$_1$__C$_2$V**	C#C'__C'V ☺	cette petite, cinq revues
Styles	J :	journalistique et conversations radio et TV du corpus ETAPE		
	F :	conversations familières du corpus NCCFr		

TABLE 1 – Récapitulatif de plan expérimental : étude du schwa en syllabe initiale de mots polysyllabiques en divers contextes consonantiques (classes C'={plosives, nasales, fricatives} ; L=[l ,ʁ]) ignorant ou non le mot précédent.

3 Résultats et discussion

3.1 Quantification des contextes avec schwa

La table 2 présente le nombre d'occurrences dans chacune des conditions examinées. Alors que les mots contenant un schwa sont assez fréquents (12-15% des mots-tokens du corpus), il n'y a que 2-4 % des mots-tokens où le schwa est situé dans des mots polysyllabiques, dont plus que la moitié se trouvent dans la syllabe initiale du mot. Parmi ceux-ci, les sous-ensembles de mots de Forme$_{2C}$ et Forme$_{3C}$ sont indiqués dans les trois dernières lignes du tableau.

Corpus :	J (Etape)	F (NCCFr)	J (Etape)	F (NCCFr)	
	Tokens		Types		Exemples
Nombre de mots dans les corpus	562041	441391	14730	11551	
dont :					
mot avec schwa (C_)	84252	54145	2195	1210	ce,demain,simplement
polysyllabique avec schwa (C_C)	21950	10958	2052	1185	demain,simplement
avec schwa en syllabe initial (#C_1_C_2)	11508	6464	906	533	demain, petite,secret
Forme$_{2C}$ (#C_1_C_2V)	7433	5003	631	383	demain, petite,
Forme$_{3C}$ (#C_1_C_2C_3V)	2887	902	214	111	secret,retraite

TABLE 2 – Nombre d'occurrences (tokens) et de mots (types) pour les deux corpus J et F et pour les sous-ensembles de mots incluant un schwa en position : C_, C_C, #C_1_C_2, #C_1_C_2V et #C_1_C_2C_3V.

3.2 Contraintes phonotactiques lexicales

3.2.1 Réalisation du schwa en syllabe initiale de mot polysyllabique

Dans cette partie, nous discutons du comportement de schwa pour les Forme$_{2C}$: #C_1_C_2V) et Forme$_{3C}$ (#C_1_C_2C_3V) dans les deux corpus sans considérer le contexte pré-frontière. Comme illustré dans la figure 1, nous observons plus de chute de schwa au sein du mot pour la Forme$_{2C}$ que pour la Forme$_{3C}$ aussi bien dans le corpus J (journalistique) que dans le corpus F (familier). La chute de schwa est donc moins fréquente si elle provoque une suite de trois consonnes qu'une suite de deux consonnes, comme prédit par la règle des 3 consonnes de Grammont.

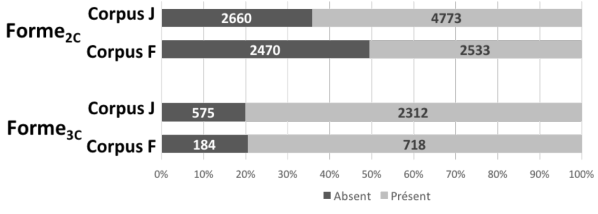


FIGURE 1 – Réalisation du schwa dans les mots de type Forme$_{2C}$ (#C_1_C_2V) et Forme$_{3C}$ (#C_1_C_2C_3V) en fonction des corpus (J et F). « Présent » indique qu'un segment schwa a été aligné ; « Absent » indique qu'un segment schwa n'a pas été aligné.

D'autre part, comme on pouvait s'y attendre, la chute de schwa est globalement plus fréquente dans le corpus de parole conversationnelle (F) que dans le corpus journalistique (J) (Brognaux *et al.*, 2014). La comparaison de ces deux facteurs influençant la réalisation ou non de schwa révèle une interaction intéressante : la différence de réalisation ou non du schwa entre les corpus J et F est plus grande pour la Forme$_{2C}$ que pour la Forme$_{3C}$. En effet, le schwa chute plus fréquemment dans le corpus de parole conversationnelle F (49%) que dans le corpus plus formel J (36%) pour les mots de Forme$_{2C}$, c'est à dire des mots où la loi des trois consonnes ne s'applique pas. Par contre, quand on regarde les mots de Forme$_{3C}$ les restrictions en termes de nombre de consonnes agissent de la même façon indépendamment du style de parole (F et J : 20%).

3.2.2 Réalisation du schwa en tenant compte de la nature des consonnes

Parmi les mots des Forme$_{2C}$ et Forme$_{3C}$, seules certaines combinaisons sont communes aux deux corpus : pour la Forme$_{2C}$: #L_LV (ex. religieux), #L_C'V (ex. recherche), #C'_LV (ex. cela) et

#C'_C'V (ex. cheveux) ; et pour la Forme$_{3C}$: #L_LGV (ex. relief), #L_C'LV (ex. regretter), #L_C'GV (ex. rejoindre), #C'_LGV (ex. celui), #C'_C'LV (ex. secret) et #C'_C'GV (ex. depuis).

Comme on peut le voir sur les tableaux 3 et 4, certaines combinaisons de consonnes apparaissent plus souvent que les autres. Comme indiqué dans la partie méthode, nous ne retiendrons pour la suite de l'analyse que les suites ayant plus de 500 occurrences dans les deux corpus. L'analyse se trouve donc restreinte au Formes$_{2C}$.

C_2 / $\#C_1$	C'		L		G	
	J	F	J	F	J	F
#C'	5148	3843	1582	796		
#L	4516	1738	262	87		
#G						

TABLE 3 – Nombre de mots de Forme$_{2C}$ en fonction des corpus J (ETAPE) et F (NCCFr) et de la natures des consonnes dans la séquence (C' : non- approximante, L : liquide, G : Glide).

C_2C_3 / $\#C_1$	C'C'		C'L		C'G		LC'		LL		LG		GC'		GL		GG	
	J	F	J	F	J	F	J	F	J	F	J	F	J	F	J	F	J	F
#C'			315	119	829	349					215	97						
#L			1339	284	301	123					2	10						
#G																		

TABLE 4 – Nombre de mots de Forme$_{3C}$ en fonction des corpus J (ETAPE) et F (NCCFr)et de la natures des consonnes dans la séquence (C' : non- approximante, L : liquide, G : Glide).

La figure 2 présente les taux (et nombre) de chute de schwa dans la Forme$_{2C}$ des corpus (J et F) en fonction de la nature des consonnes dans la séquence et de leur sonorité : croissante pour # C'_LV (☺), ou non-croissante pour #C'_C'V(☹) et #L'_C'V(☹)).

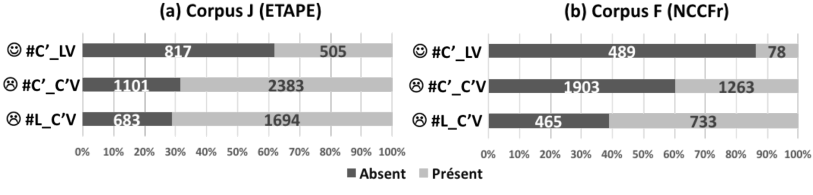


FIGURE 2 – Réalisation du schwa dans les mots de type Forme$_{2C}$ (#C_1_C_2V) en fonction des corpus (J et F) et de la nature des consonnes dans la séquence (C' : non- approximante, L : liquide).

Confirmant les données de la littérature, nous observons que la chute de schwa est plus fréquente dans les séquences où elle rend adjacentes des consonnes de sonorité croissante avec une liquide en seconde position (# C'_LV, ☺). D'ailleurs, dans ce contexte, presque tous les schwas (86%) chutent dans le corpus F.

Dans les autres contextes, si la chute de schwa est moins fréquente, elle n'en est pas moins présente. Environ 30% des schwa chutent dans les contextes #L'_C'V et #C'_C'V en parole journalistique et ces pourcentages augmentent à 39% et 60%, respectivement, dans un style plus familier du corpus F. La nature des consonnes et la sonorité croissante ou non de la séquence (telle que nous l'avons définie, voir plus loin) ne ressort donc pas comme une contrainte stricte, i.e. la chute de schwa peut mettre en contact des consonnes qui ne forme pas une suite de sonorité croissante. D'autre part, ces contraintes de sonorité semblent agir plus fortement dans une parole plus formelle (corpus J) que moins formelle (corpus F).

Toutefois, nous n'irons pas plus loin dans nos interprétations des résultats car il faut se rappeler que les contextes considérés comme non favorables à la chute de schwa, ont été définis assez grossièrement en termes de sonorité. Ainsi la catégorie des non-approximantes regroupe les plosives et les fricatives, indépendamment de leur voisement, et les nasales. Ces classes de consonnes pourraient être subdivisées plus finement en terme de sonorité (favorisant la chute de schwa dans 'semaine' vs. 'petit' par exemple). Le nombre d'occurrences présent dans les deux corpus nous a fait privilégier ce découpage grossier pour cette première analyse, mais ce point reste à affiner par la suite. De plus, nous assumons que la consonne initiale du mot (C_1) se resyllabe avec la consonne suivante (C_2) lors de la chute de schwa, or, il ne faut pas oublier que d'autres analyses comme l'extra-syllabicité (Rialland, 1986) sont possibles pour expliquer l'évitement d'une attaque à sonorité non-croissante au sein d'un mot. Sans oublier bien sûr que ces mots ne sont pas produits à l'isolé et qu'il faut prendre en compte le contexte pré-frontière précédent.

3.3 Contraintes phonotactiques post-lexicales

Dans cette partie nous examinons plus en détail deux de ces contextes pré-frontière : quand le mot précédent se termine par une voyelle (V#) ou une consonne (C#). Nous excluons ici les cas où les deux mots sont séparés par une pause ou une hésitation.

Comme on peut le voir sur la figure 3, la chute de schwa est favorisée lorsque le mot précédent se termine par une voyelle (V#), ceci dans toutes les formes ($_{2C}$ et $_{3C}$) et dans les deux corpus. Dans ce contexte, on retrouve la tendance décrite en section 3.2.1, à savoir que la chute du schwa est plus fréquente lorsqu'il provoque la succession de 2 consonnes (Forme$_{2C}$) que 3 consonnes (Forme$_{3C}$).

Lorsque le mot avec schwa est précédé d'un mot terminant par une consonne (contexte C#), on voit que l'effet du nombre de consonne agit également au niveau post-lexical. Dans les forme$_{2C}$, la consonne finale du mot précédent ajoute une troisième consonne à la suite #C_1_C_2 et on observe que le taux de délétion de schwa diminue d'environ 20% par rapport à un contexte V#. Ici encore, on note que cet évitement à former des suites de trois consonnes est indépendante du style de parole. Dans les forme$_{3C}$, dans laquelle le schwa a peu tendance à chuter, l'ajout d'une quatrième consonne pré-frontière réduit également le taux de délétion de 10% environ. Pour autant le faible nombre de cas avec chute de schwa dans ces contextes très chargés en consonnes appelle à une vérification manuelle des alignements.

Figure 4, nous croisons effet du contexte pré-frontière et nature des consonnes dans le mot avec schwa. En raison du faible nombre d'occurrences de certaines séquences, nous ne regarderons que les plus fréquentes : #C'_LV et #C'_C'V. Nous avons vu dans la section précédente que si la chute de schwa est réduite dans ces contextes lexicaux peu en accord avec le principe de sonorité, le schwa y est quand même souvent élidé, surtout dans le corpus F. L'examen plus détaillé de ces cas en fonction du mot précédent montre que la majeur partie des cas d'élision sont produit dans un contexte V#. Dans ces contextes, un possible rattachement à gauche de la consonne initiale du mot (#C_1) avec la syllabe finale précédente pourrait effectivement éviter la création d'une attaque à sonorité non-croissante. Ce rattachement à gauche est au contraire très peu probable dans les cas où le mot précédent se termine par une consonne, et l'on voit dans ces cas que l'ajout d'une consonne pré-frontière fait d'autant plus ressortir les contraintes de sonorité.

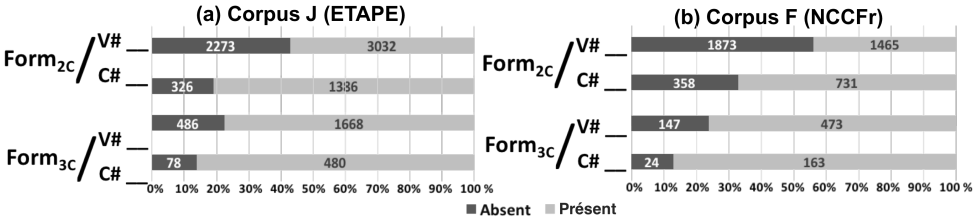


FIGURE 3 – Réalisation du schwa dans les mots de type Forme$_{2C}$ (#C$_1$_C$_2$V) et Forme$_{3C}$ (#C$_1$_C$_2$C$_3$V) en fonction des corpus (J et F) et des contextes pré-frontières (V# et C#). Autres précisions voir Figure 1.

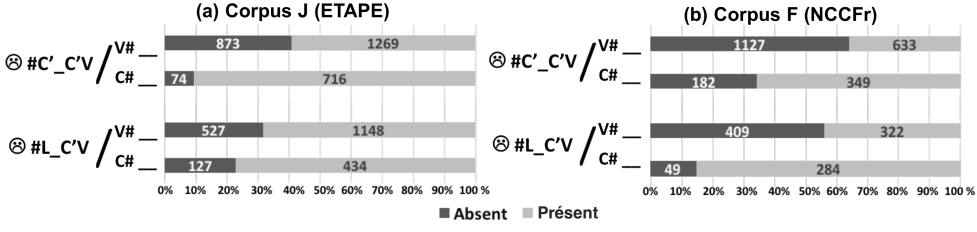


FIGURE 4 – Réalisation du schwa dans les mots de type Forme$_{2C}$ (#C$_1$_C$_2$V) en fonction des corpus (J et F), des contextes pré-frontières (V# et C#) et de la nature des consonnes dans la séquence (C' : non- approximante, L : liquide). Autres précisions voir Figure 1.

4 Conclusion

Cette étude, basée sur une grande quantité de parole à l'aide d'outils de traitement automatique, nous a permis de confirmer le rôle de la loi des trois consonnes (L3C) sur la réalisation de schwa en français. Celle-ci influence la chute de schwa tant au niveau lexical qu'au niveau post-lexical : le taux de délétion diminue dans les contextes où la chute du schwa mettrait trois consonnes ou plus à la suite, que ce soit intra ou inter mots. Nous avons également montré que cette contrainte joue aussi bien une parole plus formelle (corpus J) que moins formelle (corpus F), où le schwa chute globalement d'avantage. Les restrictions liées aux relations de sonorité entre les consonnes dans la séquence résultant de la chute de schwa ont été également mis en évidence mais contrairement à la L3C cette contrainte est moins respectée en parole familière (corpus F).

Toutefois, les cas de chute de schwa dans des séquences mettant en jeu trois ou quatre consonnes et ne respectant pas une sonorité croissante ne sont pas négligeables dans nos données, en particulier dans le corpus F. Ces cas seront à examiner plus en détail par la suite, notamment pour vérifier si les alignements sont corrects et/ou si les locuteurs n'ont pas fait chuter certaines consonnes dans la séquence pour satisfaire les contraintes liées au nombre et à la nature des consonnes autour du schwa.

Remerciements

Ce travail est financé par Investissements d'Avenir – Projet Labex EFL (ANR-10-LABX-0083) par une bourse au premier auteur. Il a été également soutenu par le projet ANR Vera.

Références

BROGNAUX S., DRUGMAN T. *et al.* (2014). Variations phonétiques : Impact de la situation de communication. *Nouveaux Cahiers de Linguistique Francaise*, **31**, 63–76.

BÜRKI A., GENDROT C., GRAVIER G., LINARÈS G. & FOUGERON C. (2008). Alignement automatique et analyse phonétique : comparaison de différents systèmes pour l'analyse du schwa. *Traitement Automatique des Langues*, **49**(3), 165–197.

CLEMENTS G. N. (1990). The role of the sonority cycle in core syllabification. **I**, 283–333.

CÔTÉ M.-H. (2000). *Consonant cluster phonotactics : a perceptual approach*. PhD thesis, Massachusetts Institute of Technology.

DELATTRE P. (1944). L'aperture et la syllabation phonétique. *The French Review*, **17**(5), 281–285.

DELATTRE P. (1966). *Studies in French and comparative phonetics*. Mouton.

DURAND J., LAKS B. & LYCHE C. (2003). Le projet" phonologie du français contemporain"(pfc). In *La tribune internationale des langues vivantes*, number 33, p. 3–10.

DURAND J., LAKS B. & LYCHE C. (2009). Le projet pfc (phonologie du français contemporain) : une source de données primaires structurées. *Phonologie, variation et accents du français. Paris : Hermès*, p. 19–61.

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 *Interspeech*, p. 1149–1152.

GAUVAIN J., ADDA G., ADDA-DECKER M., ALLAUZEN A., GENDNER V., LAMEL L. & SCHWENK H. (2005). Where Are We in Transcribing French Broadcast News ? In *Proceedings of Eurospeech-Interspeech*, p. 1665–1668, Lisbonne.

GRAMMONT M. (1894). Le patois de la franche-montagne et en particulier de damprichard (franche-comté). iv : La loi des trois consonnes. *Mémoires de la Société de linguistique de Paris*, **8**, 53–90.

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*.

LYCHE C. (1993). Quelques remarques sur le groupe ol en français. *Revue romane*, **28**(2), 195–217.

NEW B., BRYSBAERT M., VERONIS J. & PALLIER C. (2007). The use of film subtitles to estimate word frequencies. *Applied psycholinguistics*, **28**(04), 661–677.

RIALLAND A. (1986). Schwa et syllabes en français. *Studies in compensatory lengthening*, **23**, 187.

TORREIRA F., ADDA-DECKER M. & ERNESTUS M. (2010). The nijmegen corpus of casual french. *Speech Communication*, **52**(3), 201–212.

Des Réseaux de Neurones avec Mécanisme d'Attention pour la Compréhension de la Parole *

Edwin Simonnet Paul Deléglise Nathalie Camelin Yannick Estève
LIUM, Institut d'Informatique Claude Chappe Université du Maine Avenue Laennec, 72085 Le Mans, France
firstname.lastname@univ-lemans.fr

RÉSUMÉ

L'étude porte sur l'apport d'un réseau de neurones récurrent (Recurrent Neural Network - RNN) bidirectionnel encodeur/décodeur avec mécanisme d'attention pour une tâche de compréhension de la parole. Les premières expériences faites sur le corpus ATIS confirment la qualité du système RNN état de l'art utilisé pour cet article, en comparant les résultats obtenus à ceux récemment publiés dans la littérature. Des expériences supplémentaires montrent que les RNNs avec mécanisme d'attention obtiennent de meilleures performances que les RNNs récemment proposés pour la tâche d'étiquetage en concepts sémantiques. Sur le corpus MEDIA, un corpus français état de l'art pour la compréhension dédié à la réservation d'hôtel et aux informations touristiques, les expériences montrent qu'un RNN bidirectionnel atteint une f-mesure de 79,51 tandis que le même système intégrant le mécanisme d'attention permet d'atteindre une f-mesure de 80,27.

ABSTRACT

Exploring the use of Attention-Based Recurrent Neural Networks For Spoken Language Understanding

This study explores the use of a bidirectional recurrent neural network (RNN) encoder/decoder based on a mechanism of attention for a Spoken Language Understanding (SLU) task. First experiments carried on the ATIS corpus confirm the quality of the RNN baseline system used in this paper, by comparing its results on the ATIS corpus to the results recently published in the literature. Additional experiments show that RNN based on a mechanism of attention performs better than RNN architectures recently proposed for a slot filling task. On the French MEDIA corpus, a French state-of-the-art corpus for SLU dedicated to hotel reservation and tourist information, experiments show that a bidirectionnal RNN reaches a f-measure value of 79.51 while the use of a mechanism of attention allows us to reach a f-measure value of 80.27.

MOTS-CLÉS : Compréhension de la Parole, Réseaux de Neurones Récurrents, Mécanisme d'Attention, Bidirectionnel.

KEYWORDS: Spoken Language Understanding, Recurrent Neural Networks, Attention Based Mechanism, Bidirectional.

∗. Cet article présente le travail d'un commencement de thèse. Il est traduit de l'article publié en anglais à la conférence NIPS 2015 (Simonnet *et al.*, 2015)

1 Introduction

La compréhension de la parole (Spoken Language Understanding – SLU) peut être définie comme l'interprétation des signaux transportés par un signal de parole (De Mori *et al.*, 2008). Cette interprétation est habituellement assimilée à l'extraction et la représentation du *sens* contenu par les mots d'une phrase parlée.

1.1 La tâche d'étiquetage en concepts sémantiques

De nos jours, extraire le sens d'un discours reste une opération très complexe et la SLU, qui en est une application, est souvent réduite à la construction d'une représentation sémantique spécifique à la tâche.

Dans ce contexte la SLU correspond à une tâche d'étiquetage en concepts sémantiques qui est l'extraction d'une séquence de concepts sémantiques à partir d'une séquence de mots donnée (Hahn *et al.*, 2011). Dans le passé, plusieurs méthodes d'étiquetage en concepts sémantiques ont été proposées dans ce cadre. Jusqu'à il y a deux ans, les champs aléatoires conditionnels (Lafferty *et al.*, 2001)(Conditional Random Field – CRF) étaient considérés comme l'approche état de l'art (Hahn *et al.*, 2011).

1.2 Objectif

Récemment, il a été montré dans (Mesnil *et al.*, 2013, 2015) que les réseaux de neurones récurrents (Recurrent Neural Network - RNN) peuvent atteindre de meilleures performances que les CRFs dans une tâche d'étiquetage en concepts sémantiques appliquée au corpus de réservation de vol ATIS (Hemphill *et al.*, 1990). Néanmoins, (Vukotic *et al.*, 2015) a démontré que ces meilleures performances des RNNs ne se renouvellent pas sur un corpus de SLU plus complexe tel que le corpus français MEDIA (Bonneau-Maynard *et al.*, 2009) (Devillers *et al.*, 2004). En effet dans cette dernière étude, les CRFs ont obtenu des résultats significativement meilleurs que ceux des RNNs. Cela peut être expliqué par le fait que la tâche MEDIA semble plus difficile à traiter que celle d'ATIS, on remarque notamment que la taille du vocabulaire et la proportion de mots du corpus étant associé à un concept sont plus grandes dans le corpus MEDIA que dans le corpus ATIS.

Dans cet article, nous ne souhaitons pas établir si les CRFs ou les réseaux de neurones profonds (Deep Neural Network - DNN) constituent l'état de l'art courant pour la SLU. Nous sommes convaincus du potentiel des architectures DNN et nous avons pour objectif d'étudier l'utilisation d'un RNN avec mécanisme d'attention (Graves, 2013) initialement dédié à la reconnaissance d'écriture manuelle et ayant été utilisé avec succès pour la reconnaissance de la parole (Chorowski *et al.*, 2014). Étant donné que la tâche de SLU MEDIA semble plus complexe que celle d'ATIS, nous nous sommes focalisé sur le corpus MEDIA pour évaluer les conséquences de l'utilisation du mécanisme d'attention proposé dans (Bahdanau *et al.*, 2014).

1.3 Les principes généraux d'un RNN avec mécanisme d'attention

De très bonnes descriptions des principes des RNNs avec mécanisme d'attention peuvent être trouvées dans (Bahdanau *et al.*, 2014; Chorowski *et al.*, 2014, 2015). Le mécanisme d'attention fut

intuitivement conçu afin de prendre en compte la position des éléments d'entrée lors de l'encodage d'une séquence dans une approche avec un RNN encodeur-décodeur. Des poids, ré-estimés après chaque génération de sortie, sont attribués aux annotations (correspondants aux mots) en entrée. Cela permet au décodeur de décider les parties de la phrase d'entrée auxquelles prêter attention et au décodeur de ne pas avoir à encoder automatiquement toute l'information. Dans cet article, le RNN avec mécanisme d'attention s'inspire largement de l'architecture proposée dans (Bahdanau *et al.*, 2014) pour la traduction automatique, comme décrit dans la figure 1. Nous souhaitons adapter cette méthode pour la SLU en considérant le processus d'étiquetage en concepts sémantiques similaire à un problème de traduction depuis des mots (langage source) vers des concepts sémantiques (langage cible).

Cette architecture se base sur un RNN bidirectionnel comme encodeur. Ce RNN bidirectionnel calcul une annotation h_i pour chaque mot w_i de la séquence d'entrée $\{w_1, ... ,w_I\}$. Cette annotation est la concaténation des couches cachées correspondantes avant (forward) et arrière (backward) obtenues respectivement par le RNN avant et le RNN arrière constituant le RNN bidirectionnel. Chaque annotation contient le résumé d'à la fois les mots précédents et les mots suivants. Étant donné que les couches cachées des RNNs tendent à mieux représenter les entrées récentes, chaque annotation h_i se concentre sur les mots autour de w_i.

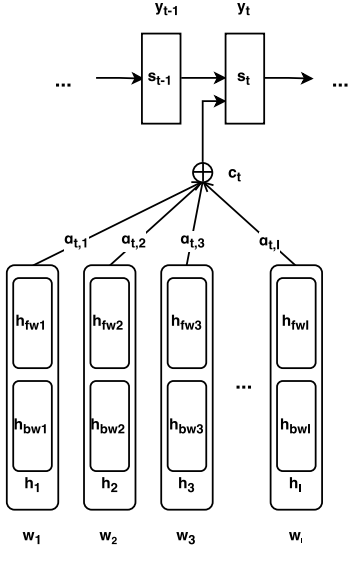


FIGURE 1 – Illustration d'un RNN avec mécanisme d'attention de (Bahdanau *et al.*, 2014)

Après avoir appliqué cet encodeur bidirectionnel, pour chaque mot à l'intérieur de la séquence d'entrée, une annotation est calculée : cette séquence d'annotations $\{h_1, ... ,h_I\}$ sera utilisée par le décodeur pour calculer un vecteur de contexte c_t. Un vecteur de contexte est recalculé après chaque émission d'une étiquette en sortie. Ce calcul prend en compte une somme pondérée de toutes les annotations calculées par l'encodeur. Cette pondération dépend de la cible en sortie courante et constitue le coeur du mécanisme d'attention : une bonne estimation des poids α_{ti} permet au décodeur de choisir les parties de la séquence d'entrée auxquelles il doit prêter attention. Ce vecteur de contexte

sera utilisé par le décodeur conjointement avec l'étiquette émise en sortie précédemment y_{t-1} et l'état courant s_t de la couche cachée du RNN afin de prendre une décision concernant l'étiquette en sortie courante y_t.

2 Implémentation

Afin de comparer les RNNs avec et sans mécanisme d'attention pour une tâche SLU, des implémentations fournies par les auteurs de (Mesnil *et al.*, 2013) et (Bahdanau *et al.*, 2014) ont été utilisées pour nos expériences. L'implémentation RNN venant de (Mesnil *et al.*, 2013) ne correspond pas à celle utilisée pour leurs expériences : seulement l'implémentation du RNN avant est disponible tandis que cette étude utilise un réseau de neurone bidirectionnel. Afin de valider notre implémentation de ce réseau de neurones bidirectionnel, des premières expériences ont été faites sur le corpus ATIS afin de comparer nos résultats avec ceux présentés dans (Mesnil *et al.*, 2013).

2.1 Implémentation d'un RNN

Notre implémentation d'un RNN se base sur (Mesnil *et al.*, 2013) et plus précisément sur l'implémentation proposée par son premier auteur (Mesnil, 2015) d'un RNN avant de type elman/jordan avec le paramètre T fixé à 1. Ce paramètre indique la récupération des T étapes temporelles précédentes depuis la couche de sortie dans un réseau de type jordan ou depuis la couche cachée dans un réseau de type elman. À partir de cette implémentation, nous utilisons un RNN de type elman [1] dont les couches cachées et couches de sorties sont calculées comme suit :

$$couche\ cachee : h(t) = sigmoid(W_x.x(t) + W_h.h(t-1) + b_h)) \ (1)$$

$$couche\ de\ sortie : s(t) = softmax(W.h(t) + b)) \ (2)$$

où $x(t)$ est le mot d'entrée du RNN (représentation continue de mot - embedding) à t et $h(t-1)$ la sortie de la couche cachée à $t-1$. Les paramètres W_x, W_h et W du RNN sont les matrices de poids, b_h et b les biais, et h_0 la couche cachée initiale de l'étape précédente pour le premier mot de la séquence pour lequel rien n'a encore été calculé. Les paramètres sont ajustés au cours des époques d'apprentissage à l'aide d'une descente de gradient effectuée sur des *mini-batches*.

La version arrière est implémentée à partir de la version avant fournie dans (Mesnil, 2015). Un RNN arrière est similaire à un avant à l'exception que la prédiction est faite du futur vers le passé. La phrase est donnée à l'envers pour simuler cet effet. W_h représente la matrice de poids entre la couche cachée de l'étape prochaine et la courante. h_0 est la couche cachée initiale de l'étape prochaine pour le dernier mot de la phrase (*i.e.* le premier mot donné au RNN). La couche de sortie est toujours calculée à l'aide de l'équation (2). La couche cachée est calculée comme suit :

$$h(t) = sigmoid(W_x.x(t) + W_h.h(t+1) + b_h)) \ (3)$$

Avec un RNN arrière acquis, le bidirectionnel peut être implémenté. Un RNN bidirectionnel effectue des prédictions prenant en compte le passé et le futur. Par conséquent, des RNNs avant et arrière déjà

1. l'opération serait la même pour un jordan excepté que la couche de sortie est réinjectée dans la couche cachée à t+1

entraînés sont utilisés conjointement. Il y a deux matrices de poids W_h : W_h_fw entre la couche cachée de l'étape précédente et la courante ; et W_h_bw entre la couche cachée de l'étape suivante et la courante. Il en va de même pour b_h_fw et b_h_bw. Finalement il n'y a pas de couches cachées initiales h_0 étant donné que ces dernières sont récupérées depuis les RNNs avant et arrière déjà entraînés. La couche cachée est alors calculée comme suit :

$$h(t) = sigmoid(W_x.x(t) + W_h\_fw.h(t-1) + b_h\_fw + W_h\_bw.h(t+1) + b_h\_bw) \quad (4)$$

Notre objectif est également d'implémenter des dépendances à long termes comme décrit dans (Mesnil *et al.*, 2013) en fournissant au réseau la somme des étapes précédentes/suivantes, c'est à dire avec un T supérieur à 1 selon l'équation :

$$h_{bd}(t) = f(W_x.x(t) + \sum_{k=1}^{T}(W_h\_bw_k.h\_bw(t+k) + b_h\_bw) + \sum_{k=1}^{T}(W_h\_fw_k.h\_fw(t-k) + b_h\_fw)) \quad (5)$$

(Mesnil, 2015) donne une implémentation avec T fixé à 1. Différentes valeurs de T ont été testées sur le corpus ATIS pour les RNNs avant et arrière afin de voir si cet ajout de contexte améliore le système, mais le RNN bidirectionnel atteint ses meilleurs résultats avec des RNNs avant et arrière ayant tous deux T=1.

Les RNNs avant et arrière utilisés pour l'entraînement ou la classification du bidirectionnel sont entraînés individuellement auparavant. Différentes manières d'entraîner ces RNNs ont été testées. D'abord le *parallel train*, qui consiste à entraîner les RNNs avant, arrière et ensuite bidirectionnel à chaque époque. Deuxièmement, l'entraînement *get best*, dans lequel l'entraînement du RNN bidirectionnel se base sur les meilleurs paramètres d'à la fois les RNNs avant et arrière déjà entraînés avant. Enfin, l'apprentissage *train best* qui combine les deux approches précédentes : à chaque époque les RNNs avant et arrière sont entraînés comme dans le *parallel train*. Ensuite le RNN bidirectionnel utilise les paramètres des dernières meilleures époques pour le RNN avant et arrière respectivement comme dans le *get best*.

Les expériences ont montré que le meilleur apprentissage est l'approche *parallel train* suivi du *train best* et enfin du *get best*. Cela peut être expliqué par le fait que le RNN bidirectionnel apprend d'avantage avec des RNNs avant et arrière qui ont une plus grande variabilité au cours des époques même s'ils ne donnent pas toujours les meilleurs résultats. L'apprentissage en est par conséquent plus diversifié. En effet l'approche *get best* utilisant des paramètres avant et arrière fixés à partir de leur meilleures époques est celle qui donne les pires résultats.

2.2 Implémentation d'un RNN bidirectionnel avec mécanisme d'attention

L'implémentation d'un RNN avec mécanisme d'attention utilisé dans notre étude est dérivée de celle utilisée dans (Bahdanau *et al.*, 2014) et disponible à `https://github.com/kyunghyuncho/GroundHog`. Cette implémentation a été créée pour une tâche de traduction automatique. Dans cette tâche, les séquences d'entrée et de sortie ont souvent des longueurs différentes. L'approche avec un RNN encodeur-décodeur est particulièrement bien adaptée pour ce cas de figure. Pour la tâche SLU en particulier, il est très important d'obtenir une correspondance entre les mots (entrées) et les concepts sémantiques (sorties) afin de pouvoir extraire la valeur du concept. Afin d'obtenir un

alignement précis, nous avons modifié le processus de décodage du RNN bidirectionnel en imposant que la séquence d'étiquettes en sortie et la séquence de mots en entrée aient la même taille. C'est la seule modification apportée à l'implémentation venant de (Bahdanau *et al.*, 2014).

3 Expériences

Afin de valider notre propre implémentation du réseau de neurones bidirectionnel et comparer nos résultats avec ceux présentés dans (Mesnil *et al.*, 2013), nous avons mené des premières expériences sur le corpus ATIS. Puis, une fois notre implémentation validée, nous comparons les approches RNN avec et sans mécanisme d'attention sur le corpus MEDIA.

3.1 Validation de l'implémentation d'un RNN sur le corpus ATIS

Le corpus utilisé par (Mesnil *et al.*, 2013) est le corpus ATIS (Airline Travel Information System) spécialisé dans les requêtes de réservation de billets d'avion. Il est composé de 4978/893 (apprentissage/test) phrases annotées selon 128 étiquettes sémantiques. Le corpus d'apprentissage est divisé comme suit : 80% pour l'apprentissage 20% pour la validation.

Afin d'aider le classifieur à délimiter les séquences de mots ayant la même étiquette, une méthode courante consiste à rajouter un suffixe $B/I/O$ aux étiquettes sémantiques, respectivement pour le début (*Beginning*), l'intérieur (*Inside*) et l'extérieur (*Outside*) d'une séquence. Seuls les suffixes B et I sont utilisés ici. O est représenté par l'étiquette $NULL$ qui est associée aux mots ne portant aucune information sémantique selon la tâche spécifique.

L'évaluation est faite en calculant la f-mesure qui utilise les métriques de rappel et de précision. On obtient un score prenant en compte la présence ou l'absence d'un concept dans une phrase sans notion de séquentialité.

$$f - mesure = \frac{2(precision.rappel)}{precision + rappel}$$

Dans (Mesnil *et al.*, 2013) précision et rappel sont définis [2] comme suit :

$$rappel = \frac{nombre\ de\ segments\ corrects}{nombre\ de\ segments\ dans\ la\ reference}$$

$$precision = \frac{nombre\ de\ segments\ corrects}{nombre\ de\ segments\ dans\ l'\ hypothese}$$

Un segment de concept est correct s'il commence et finit avec les mêmes mots pour l'hypothèse et la référence. La f-mesure est maximisée sur le corpus de validation durant le processus d'apprentissage.

Le tableau 1 présente les résultats obtenus par (Mesnil *et al.*, 2013) et ceux par notre implémentation en utilisant les hyper-paramètres suivants : nombre d'époques=100 ; fenêtre=5 ; nombre d'unités dans la couche cachée=200 ; dimension d'embedding=50.

2. Calculés à l'aide du script conlleval.pl fourni par (Mesnil, 2015)

Expérience	Architecture	Type	f-mesure
[Mesnil et al. 2013]	Jordan	bidirectional	93,98
RNN baseline	Elman	bidirectional	94,13

TABLE 1 – Comparaison entre les performances du RNN présenté dans (Mesnil *et al.*, 2013) et notre implémentation d'un RNN état de l'art sur le corpus ATIS.

Comme montré dans le tableau 1, notre système RNN bidirectionnel état de l'art atteint des résultats similaires à ceux du RNN bidirectionnel présenté dans (Mesnil *et al.*, 2013). Cela valide notre implémentation et nous permet d'étudier l'utilisation d'un RNN avec mécanisme d'attention pour une tâche SLU d'étiquetage en concepts sémantiques plus complexe.

3.2 Performances d'un RNN avec mécanisme d'attention sur le corpus MEDIA

Le corpus MEDIA (Bonneau-Maynard *et al.*, 2009) est un corpus de dialogue état de l'art français. Il contient 1257 dialogues entre des utilisateurs et un système simulé (protocole Wizard of Oz) dans le domaine de la réservation d'hôtel et des informations touristiques. Seuls les tours de parole des utilisateurs sont utilisés pour l'apprentissage et la classification. Cet ensemble de tours est divisé en trois sous-corpus : l'ensemble APPRENTISSAGE qui contient 17,6k énoncés, l'ensemble DEV qui contient 1,3k énoncés et enfin l'ensemble TEST qui est composé de 3,5k énoncés.

Chaque énoncé a été manuellement transcrit et annoté en se basant sur 74 étiquettes de concept allant de simples réponses (*e.g.* le mot "oui" est associé au concept *reponse*) à des requêtes spécifiques à la tâche (*e.g.* les mots "avec baignoire" sont associés au concept *equipement_chambre*). Une annotation plus riche est disponible dans MEDIA incluant les modes, les spécifieurs et les valeurs mais pour commencer nous choisissons d'évaluer seulement les étiquettes des concepts sémantiques.

Dans MEDIA, le but du dialogue pour l'utilisateur est d'obtenir des informations qui sont stockées dans une base de données. Par conséquent, les noms de rues, de villes ou d'hôtels, les listes d'équipement de chambre, les types de nourriture, *etc.* sont connus. De plus, des mots plus généraux représentant les nombres, les jours, les mois sont également connus. Tous ces mots (spécifiques à la tâche SLU ou généraux) ont été rassemblés dans un lexique sémantique qui permet d'associer un mot à une classe sémantique.

L'énoncé d'un utilisateur est représenté par une séquence de mots et de classes sémantiques. Si elle existe, le mot est substitué par sa classe sémantique. Un exemple est disponible dans le tableau 2. Comme pour ATIS, les suffixes B/I sont associés aux étiquettes de concepts.

Mots	j'aimerais réserver un hôtel pour les **trois premiers** jours de **Mai** à **Marseille**.
Mots+Class. Sem.	j'aimerais réserver un hôtel pour les **UNIT ORDINAL** jours de **MOIS** à **VILLE**.

TABLE 2 – Représentation d'un énoncé utilisateur par mots et par mots+classes sémantiques.

Le tableau 3 montre les performances mesurées en terme de f-mesure des différentes architectures de RNN sur le corpus MEDIA.

Architecture	f-mesure
RNN avant	74,04
RNN arrière	77,42
RNN bidirectionnel	79,51
Mécanisme d'attention	80,27
RNN encodeur-décodeur sans mécanisme d'attention	38,25

TABLE 3 – Résultats sur MEDIA avec classes sémantiques

Comme prévu, les résultats ne sont pas aussi bon que sur le corpus ATIS. L'architecture RNN bidirectionnel obtient de meilleurs résultats en comparaison avec un RNN arrière ou avant. Cela confirme l'utilité d'utiliser des informations du contexte passé et futur ensemble.

De plus, les résultats présentés dans le tableau 3 montrent que l'encodeur RNN bidirectionnel avec mécanisme d'attention est plus performant qu'un RNN bidirectionnel classique.

Enfin, il est montré qu'un encodeur-décodeur RNN obtient de très faibles performances sans mécanisme d'attention lors de la production d'une séquence d'étiquettes en sortie ayant la même longueur que la séquence de mots en entrée.

4 Conclusion

Cette étude a pour but d'examiner l'utilisation d'un RNN bidirectionnel avec mécanisme d'attention pour la tâche SLU d'étiquetage en concepts sémantiques. Nos expériences montrent que cette architecture atteint de meilleurs résultats qu'une approche plus classique avec un RNN bidirectionnel sur un corpus SLU complexe. Cette approche classique d'un RNN bidirectionnel a été introduite et présentée comme l'approche état de l'art pour la SLU il y a 2 ans par (Mesnil *et al.*, 2013) sur le corpus ATIS. Même si (Vukotic *et al.*, 2015) a montré que les CRFs obtiennent toujours de meilleurs résultats que ce RNN bidirectionnel sur des données plus complexes comme le corpus MEDIA, nos résultats montrent que des approches prometteuses comme le mécanisme d'attention peuvent toujours améliorer les RNN pour la SLU.

Remerciements

Nous remercions l'agence ANR pour son financement à travers CHIST-ERA ERA-Net JOKER sous le numéro de contrat ANR-13-CHR2-0003-05.

De plus, les auteurs remercient Sahar Ghannay pour son aide constructive au cours de l'écriture de cet article.

Références

BAHDANAU D., CHO K. & BENGIO Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*.

BONNEAU-MAYNARD H., QUIGNARD M. & DENIS A. (2009). Media : a semantically annotated corpus of task oriented dialogs in french. *Language Resources and Evaluation*, **43**(4), 329–354.

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*.

CHOROWSKI J., BAHDANAU D., SERDYUK D., CHO K. & BENGIO Y. (2015). Attention-based models for speech recognition. *arXiv preprint arXiv :1506.07503*.

DE MORI R., BECHET F., HAKKANI-TÜR D., MCTEAR M., RICCARDI G. & TUR G. (2008). Spoken language understanding. *Signal Processing Magazine, IEEE*, **25**(3), 50–58.

DEVILLERS L., MAYNARD H., ROSSET S., PAROUBEK P., MCTAIT K., MOSTEFA D., CHOUKRI K., CHARNAY L., BOUSQUET C., VIGOUROUX N., BÉCHET F., ROMARY L., ANTOINE J., VILLANEAU J., VERGNES M. & GOULIAN J. (2004). The french media/evalda project : the evaluation of the understanding capability of spoken language dialogue systems. In *LREC*.

GRAVES A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv :1308.0850*.

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. *Audio, Speech, and Language Processing, IEEE Transactions on*, **19**(6), 1569–1583.

HEMPHILL C. T., GODFREY J. J. & DODDINGTON G. R. (1990). The atis spoken language systems pilot corpus. In *Proceedings of the DARPA speech and natural language workshop*, p. 96–101.

LAFFERTY J. D., MCCALLUM J. D. & PEREIRA F. C. N. (2001). Conditional random fields : probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML-2001)*, San Francisco, CA, USA.

MESNIL G. (2015). Recurrent Neural Networks with Word Embeddings DeepLearning 0.1 documentation. `http://www.deeplearning.net/tutorial/rnnslu.html#rnnslu`.

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. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, **23**(3), 530–539.

MESNIL G., HE X., DENG L. & BENGIO Y. (2013). Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *INTERSPEECH*, p. 3771–3775.

SIMONNET E., DELÉGLISE P., CAMELIN N. & ESTÈVE Y. (2015). Exploring the use of attention-based recurrent neural networks for spoken language understanding. In *NIPS*.

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*.

Un Sous-espace Thématique Latent pour la Compréhension du Langage Parlé

Mohamed Bouaziz[1, 2] Mohamed Morchid[1] Pierre-Michel Bousquet[1]
Richard Dufour[1] Killian Janod[1] Waad Ben Kheder[1] Georges Linarès[1]
(1) LIA, 339 Chemin des Meinajaries, 84140 Avignon, France
(2) EDD, 28 Boulevard de Port-Royal, 75005 Paris, France
mohamed.bouaziz@alumni.univ-avignon.fr, {prénom.nom}@univ-avignon.fr

RÉSUMÉ

Les applications de compréhension du langage parlé sont moins performantes si les documents transcrits automatiquement contiennent un taux d'erreur-mot élevé. Des solutions récentes proposent de projeter ces transcriptions dans un espace de thèmes, comme par exemple l'allocation latente de Dirichlet (LDA), la LDA supervisée ainsi que le modèle *author-topic* (AT). Une représentation compacte originale, appelée *c*-vector, a été récemment introduite afin de surmonter la difficulté liée au choix de la taille de ces espaces thématiques. Cette représentation améliore la robustesse aux erreurs de transcription, en compactant les différentes représentations LDA d'un document parlé dans un espace réduit. Le défaut majeur de cette méthode est le nombre élevé de sous-tâches nécessaires à la construction de l'espace *c*-vector. Cet article propose de corriger ce défaut en utilisant un cadre original fondé sur un espace de caractéristiques robustes de faible dimension provenant d'un ensemble de modèles *AT* considérant à la fois le contenu du dialogue parlé (les mots) et la classe du document. Les expérimentations, conduites sur le corpus DECODA, montrent que la représentation proposée permet un gain de plus de 2.5 points en termes de conversations correctement classifiées.

ABSTRACT

A Latent Topic-based Subspace for Spoken Language Understanding.

Performance of spoken language understanding applications declines when spoken documents automatically transcribed contain high Word Error Rates (WER). Recent solutions propose to map these automatic transcriptions in topic-based representations, such as Latent Dirichlet Allocation (LDA), supervised LDA and author-topic (AT) models. An original compact representation, called *c*-vector, has recently been introduced to walk around the tricky choice of the number of latent topics in these topic-based representations. This representation increases the robustness of document classification with respect to transcription errors by compacting different LDA representations of a same speech document in a reduced space and then compensate most of the noise of the document representation. The main drawback of this method is the number of sub-tasks needed to build the *c*-vector space. This paper proposes to tackle this drawback using an original framework in a robust low dimensional space of features from a set of AT models, considering not only the dialogue content (words), but also the class related to the document. Experiments, conducted on the DECODA corpus, show that the original LTS representation allows a gain of more than 2.5 points in terms of correctly labeled conversations.

MOTS-CLÉS : Modèle Author-Topic, Analyse factorielle, c-vecteur, Classification de documents.

KEYWORDS: Author-Topic model, Factor analysis, c-vector, Document classification.

1 Introduction

Les performances des applications de compréhension du langage diminuent lorsque les transcriptions automatiques des documents parlés contiennent un grand nombre d'erreurs. L'étape de reconnaissance de la parole doit faire face à des difficultés d'origines (conditions de prise du son, bruits environnementaux, nombreuses disfluences...).

Des travaux récents portant sur l'analyse des conversations parlées, l'analyse de la parole, l'identification du sujet et la segmentation, peuvent être trouvés dans (Eisenstein & Barzilay, 2008; Lagus & Kuusisto, 2002; Tur & De Mori, 2011; Hazen, 2011; Melamed & Gilbert, 2011) et (Purver, 2011). Une manière efficace pour améliorer la robustesse des Systèmes de Reconnaissance Automatique de la Parole (SRAP) est de projeter les conversations dans un espace abstrait permettant une classification efficace de ces dialogues dans un espace caché robuste. De nombreux espaces thématiques non supervisés ont été proposés afin de représenter efficacement le contenu d'un dialogue comme l'allocation latente de Dirichlet (*Latent Dirichlet Allocation* ou LDA) (Blei *et al.*, 2003) et les modèles *author-topic* (AT) (Rosen-Zvi *et al.*, 2004). (Morchid *et al.*, 2014a) et (Morchid *et al.*, 2014c) ont respectivement surmonté deux difficultés :

— réussir à choisir le nombre de thèmes optimal de l'espace thématique en utilisant plusieurs représentations latentes obtenues en variant la taille de l'espace LDA puis en compactant ces représentations à l'aide de l'analyse factorielle (Dehak *et al.*, 2011) (diverses sous-tâches sont requises),

— contraindre les méthodes de recherche de thèmes latents, qui sont non supervisées pour la plupart, à se rapprocher des thématiques recherchées dans la tâche. Dans ce but, le modèle *author-topic* a été défini pour prendre en considération à la fois le contenu des documents (*i.e.* les mots), la thématique du document, lorsque celle-ci est connue, mais aussi la distribution des mots sachant la thématique qui est modélisée par une relation latente.

Cet article propose tout d'abord de traiter conjointement les deux problèmes ci-dessus en apprenant un grand nombre de modèles AT et, ensuite, en extrayant de ces représentations des vecteurs compacts au moyen de l'analyse factorielle. Cette approche nécessite de nombreux pré-traitements ou projections (réseau de neurones profond (May *et al.*, 2015), UBM-GMM, normalisation...), les gains les plus importants étant observés sur des données très bruitées (Bousquet *et al.*, 2011). Néanmoins, notons que des représentations telles que les modèle AT ne contiennent qu'une faible portion de bruit due à la taille réduite de cette représentation comparativement à un espace LDA classique. Ainsi, ce papier propose de considérer les différents modèles AT comme un sous-espace de caractéristiques commun et de compacter ces différentes représentations afin d'extraire directement un vecteur robuste de faible dimension (super-vecteur).

La suite de ce papier est organisée comme suit. Les approches proposées sont décrites dans la section 2. La section 3 présente le protocole expérimental et expose les résultats. Enfin, la section 5 conclut ce travail et offre quelques perspectives.

2 Approche proposée

L'approche originale proposée, appelée "sous-espace thématique latent" (*Latent Topic-based Subspace* ou LTS), est comparée avec le représentation c-vecteur (Morchid *et al.*, 2014a). Les deux approches apprennent un ensemble d'espaces AT détaillés dans la section 2.1, puis projettent les différents

documents dans des espaces thématiques. Enfin, une compression de ces représentations est effectuée avec l'analyse factorielle pour la représentation fondée sur le c-vecteur, et avec la Décomposition en Éléments Propres (*Eigenvalues Decomposition* ou EVD) pour le LTS. La section 2.2 décrit l'approche c-vecteur illustrée dans la figure 1-(a)-(b)-(c), l'approche proposée (LTS) étant présentée dans la section 2.3. En ce qui concerne la technique LTS, les divers espaces thématiques sont considérés comme des sous-espaces latents homogènes, ce qui évite de projeter les documents dans un GMM. En outre, les super-vecteurs qui composent le LTS sont compressés à l'aide d'une EVD simple afin d'extraire une représentation robuste du document. Ces méthodes sont décrites dans la section suivante.

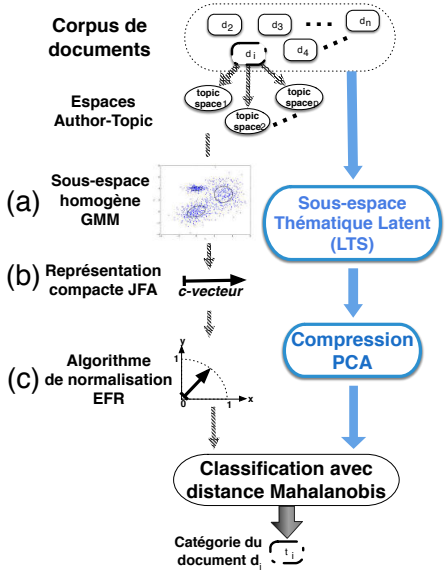


FIGURE 1 – Le sous-espace JFA + GMM ((a)-(c)) et la compression LTS (en bleu).

2.1 Modèle *Author-topic* (AT)

Le modèle AT encode à la fois le contenu du document (la distribution des mots) et les auteurs (distribution des auteurs). Dans le cadre de notre application, un document d est une conversation entre humains agent/utilisateur. L'agent doit annoter cette conversation avec un des 8 thèmes définis en amont. Les thèmes correspondent ici aux "auteurs" dans les modèles AT. Chaque dialogue d est composé d'un ensemble de mots w et un thème a. Dans ce modèle, chaque auteur est associé à une distribution sur les thèmes latents (θ), choisie à l'aide d'une loi de Dirichlet ($\overrightarrow{\alpha}$) et à une mixture pondérée afin de sélectionner un thème z. Un mot est alors généré selon la distribution ϕ correspondant au thème z. Cette distribution ϕ est déterminée à partir d'un Dirichlet ($\overrightarrow{\beta}$) Par conséquent, ce modèle permet d'encoder les dépendances statistiques entre le contenu du dialogue (les mots w) et le label (thème a) à partir de la distribution du thème latent z dans le dialogue. L'échantillonnage de Gibbs nous permet d'estimer les paramètres du modèle AT afin de représenter un dialogue non observé d avec le r^{eme} espace AT de taille T et d'obtenir un vecteur de paramètres $V_d^{a_k^r} = P(a_k|d)$ relatif à la

représentation du thème d'un dialogue non observé d selon le r^{eme} espace AT Δ_r^n de taille T. Le k^{eme} $(1 \leq k \leq A)$ paramètre est :

$$V_{d,r}^{a_k^r} = \sum_{i=1}^{N_d} \sum_{j=1}^{T} \theta_{j,a_k}^r \phi_{j,i}^r \quad (1)$$

où A est le nombre de thèmes ; $\theta_{j,a_k}^r = P(a_k|z_j^r)$ est la probabilité du thème a_k à être généré par le thème z_j^r dans le r^{eme} espace de thèmes de taille T. $\phi_{j,i}^r = P(w_i|z_j^r)$ est la probabilité du mot w_i (N_d est la taille du vocabulaire de d) à être généré par le thème z_j^r.

2.2 Représentation c-vecteur

Cette approche, initialement proposée dans (Morchid *et al.*, 2014b), utilise les i-vecteurs pour modéliser la représentation du dialogue à travers chaque espace AT dans un espace homogène. Ces segments courts sont considérés comme des unités de représentation sémantique. Dans notre modèle, le super-vecteur m d'un dialogue d sachant un espace thématique r, est issu de la concaténation des moyennes de chacune des Gaussiennes composant le modèle de GMM (UBM) :

$$\mathbf{m}_{(d,r)} = m + \mathbf{T}\mathbf{x}_{(d,r)} \quad (2)$$

où $\mathbf{x}_{(d,r)}$ contient les coordonnées de la représentation AT du dialogue dans l'espace de variabilité totale réduite ; m est le super-vecteur moyen de l'UBM. $\mathbf{T}$ est la *matrice de variabilité totale* de faible dimension $(MD \times R)$, où M est le nombre de Gaussiennes dans l'UBM et D est la taille des caractéristiques. La représentation c-vecteur souffre de 3 problèmes : (i) les c-vecteurs x de l'équation 2 doivent être distribués au sein de la distribution normale $\mathcal{N}(0, I)$, (ii) leur effet "radial" doit être supprimé et (iii) l'espace factoriel de plein rang doit être utilisé pour appliquer des transformations discriminantes. Une solution à ces 3 problèmes (*Eigen Factor Radial* (EFR)) a été développée dans (Bousquet *et al.*, 2011) en standardisant les c-vecteurs comme décrit dans la figure 2.

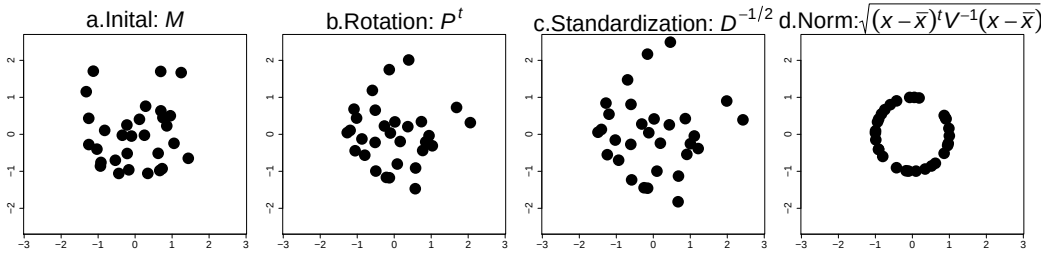


FIGURE 2 – Effet de la standardisation des c-vecteurs à l'aide de l'algorithme EFR.

2.3 Sous-espace Thématique Latent (LTS)

La représentation c-vecteur doit projeter les dialogues dans un UBM-GMM afin d'obtenir un super-vecteur de très grande dimension (la taille de la représentation thématique mutipliée par le nombre de Gaussiennes dans l'UBM). Le LTS est composé d'un ensemble d'espaces latents et considère chacun de ces derniers comme un sous-domaine sur lequel est projeté chaque document. Ainsi, toutes les

représentations thématiques d'un document partagent une structure latente commune. Ces paramètres latents partagés définissent le sous-espace thématique latent. Chaque super-vecteur s_d, représentant un document d d'un ensemble de documents de taille N, est partiellement associé à un petit sous-ensemble de caractéristiques latentes et le résidu de cette représentation du document est projeté dans un espace de caractéristiques global partagé par toutes les représentations définissant le sous-espace latent. Le super-vecteur $\mathbf{s}_d$ relatif à un dialogue d, est obtenu en concaténant les représentations AT $V_{d,r}^{a_k^r}$ pour tous les espaces thématiques r. Ainsi, la matrice $\mathbf{S} = [s_0, \ldots, s_d, \ldots, s_N]$ des super-vecteurs représente les documents dans le LTS. Cette matrice est ensuite compressée au moyen d'une EVD aboutissant à une représentation compacte $\mathbf{h}_d$ dans un espace de faible dimension dont la taille dépend du nombre des valeurs propres e :

$$\mathbf{S} = \mathbf{P}\Delta\mathbf{V}^T \tag{3}$$

où $\mathbf{P}$ est la matrice des vecteurs singuliers à gauche de taille $MD \times N$, $\mathbf{V}$ est la matrice des vecteurs singuliers à droite de taille $N \times N$ ($N << MD$) et $\mathbf{\Delta}$ est la matrice diagonale des valeurs singulières. N est le rang de la matrice $\mathbf{S}$. Plus d'informations sur l'EVD sont disponibles dans (Abdi & Williams, 2010) et (Golub & Reinsch, 1970). La représentation compacte $\mathbf{h}_{(d,e)}$, de taille e (nombre des valeurs propres), relative à un super-vecteur $\mathbf{s}_d$ de $\mathbf{S}$, est définie comme suit :

$$\mathbf{h}_{(d,e)} = (\mathbf{s}_d - \overline{s}) \, . \mathbf{V}_e^T \tag{4}$$

où $\mathbf{V}_e$ est la matrice réduite de vecteurs propres relative aux e plus grandes valeurs propres présentes dans la matrice diagonale Δ et $\overline{s}$ est la moyenne de tous les super-vecteurs des documents. En outre, cette représentation compacte d'un document ne nécessite ni l'apprentissage d'un espace commun (comme un UBM-GMM), les espaces thématiques étant un espace de caractéristiques homogène (figure 1-(a) et (b)), ni la normalisation du super-vecteur avec, par exemple, l'algorithme EFR (figure 1-(c)).

3 Protocole Expérimental

L'efficacité de la représentation compacte proposée dans le LTS est évaluée dans le cadre du corpus DECODA (Bechet *et al.*, 2012). Ce corpus comporte environ 74 heures d'enregistrements audio relatifs à 1 514 conversations téléphoniques manuellement annotées en 8 ($A = 8$) thèmes : *Itinéraire, Objets trouvés, Horaire, Carte de transport, État du trafic, Prix du ticket, Infractions, Offres spéciales*. 740 conversations servent pour la phase d'apprentissage, 175 pour le développement et 327 pour le test.

La transcription des dialogues a été effectuée par le SRAP Speeral du LIA (Linarès *et al.*, 2007). Les paramètres du modèle acoustique ont été estimés à partir de 150 heures de conversations téléphoniques. Le vocabulaire contient 5 782 mots. Un modèle de langage 3-grammes a été obtenu en adaptant un ML basique sur les transcriptions du corpus d'apprentissage. Ce système atteint un taux d'erreur de mots de 33.8% sur le corpus d'apprentissage, 45.2% sur le corpus de développement, et 49.5% sur celui de test. Ces taux élevés sont dus aux disfluences et aux environnements acoustiques défavorables (par exemple, un appel depuis une rue bruyante avec un téléphone portable).

Une approche de classification fondée sur la distance de Mahalanobis (Morchid *et al.*, 2014a) est utilisée pour trouver le thème principal d'un dialogue. Cette approche probabiliste ignore le processus par lequel les vecteurs ont été extraits. Une fois qu'un vecteur compact a été obtenu à partir d'un

document, le mécanisme de sa représentation est ignoré et il est considéré comme une observation provenant d'un modèle probabiliste génératif. La métrique de Mahalanobis affecte un document d au thème le plus probable C. Sachant un ensemble de documents d'apprentissage, soit $\mathbf{W}$ la matrice de covariance intra-document définie par :

$$\mathbf{W} = \sum_{k=1}^{K} \frac{n_t}{N} \mathbf{W_k} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=0}^{n_t} \left(x_i^k - \overline{x_k}\right) \left(x_i^k - \overline{x_k}\right)^t \tag{5}$$

où $\mathbf{W_k}$ est la matrice de covariance du k^{eme} thème C_k, n_t le nombre de phrases dans le thème C_k, N est le nombre total de documents, et $\overline{x_k}$ la moyenne de tous les documents x_i^k de C_k. Les documents ne contribuent pas de la même manière dans la covariance. C'est pour cette raison que le terme $\frac{n_t}{N}$ est introduit dans l'équation 5. En assumant l'homoscédasticité (égalité entre les classes de covariances) et la Gaussianité de la densité conditionnelle, une nouvelle observation x dans les données de test peut être affectée au thème le plus probable $C_{k_{\text{Bayes}}}$ en utilisant le classifieur s'appuyant sur la loi de décision de Bayes :

$$C_{k_{\text{Bayes}}} = \arg\max_k \left\{ -\frac{1}{2} \left(x - \overline{x_k}\right)^t \mathbf{W}^{-1} \left(x - \overline{x_k}\right) + a_k \right\} \tag{6}$$

où $a_k = \log\left(P(C_k)\right)$. Il est à noter que, avec ces hypothèses, l'approche bayésienne est similaire à l'approche géométrique de Fisher. x est affectée à la classe de la plus proche moyenne, selon la métrique de Mahalanobis (Xing *et al.*, 2002) de $\mathbf{W}^{-1}$:

$$C_{k_{\text{Bayes}}} = \arg\max_k \left\{ -\frac{1}{2} ||x - \overline{x_k}||^2_{\mathbf{W}^{-1}} + a_k \right\} \tag{7}$$

4 Expériences et résultats

La section 4.1 présente les résultats obtenus avec deux représentations fondées sur les modèles AT et la technique c-vecteur. Ensuite, la proposition compacte originale s'appuyant sur un LTS est comparée avec la représentation c-vecteur dans la section 4.2.

4.1 Impact de la compression c-vecteur

Les expériences sont conduites en utilisant 500 espaces AT en faisant varier le nombre de thèmes de 2 à 505 par pas de 1. Une solution classique consiste à chercher l'espace thématique qui atteint la meilleure performance. La figure 3 présente les performances de classification de thèmes obtenues sur le corpus de développement (figure 3) et sur le corpus de test (figure 3-(b)) en utilisant diverses configurations de la représentation AT (*baseline*).

Premièrement, nous pouvons constater que la méthode baseline atteint une précision de 86,3 % et 83,8 % respectivement sur les données de développement et de test. Cependant, nous remarquons que la performance de classification est plutôt instable et peut varier complètement d'un espace thématique à l'autre. L'écart entre les bornes inférieures et supérieures des résultats de classification est aussi important (21,2 points). Par conséquent, il est crucial de trouver la meilleure configuration d'espace AT pour cette tâche de classification. En utilisant les paramètres du point de fonctionnement

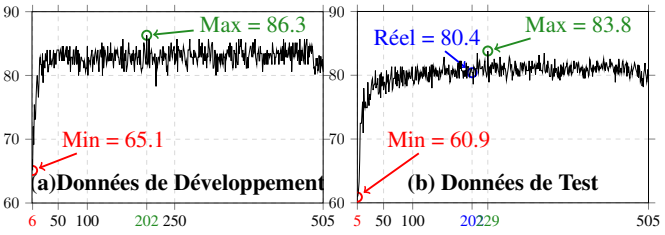


FIGURE 3 – La précision de la classification en thèmes (%) en utilisant diverses représentations thématiques sur les corpus de développement et de test avec différentes configurations expérimentales. L'axe X représente le nombre n de classes composant l'espace thématique ($5 \leq n \leq 505$).

($n = 202$ thèmes) estimé sur le corpus de développement, la classification atteint une précision de 80,4 %.

Le tableau 1 présente la méthode c-vecteur associée à l'algorithme de normalisation EFR (Morchid *et al.*, 2014a). Nous pouvons premièrement remarquer que cette représentation compacte permet de surpasser la performance de la configuration d'espace AT optimale avec un gain de 1,9 points sur les données de test. L'incohérence de la performance de classification n'est pas observée avec cette approche (comme déjà mentionné dans (Morchid *et al.*, 2014a)). En effet, la configuration qui a obtenu la meilleure précision sur les données de développement est aussi optimale pour les données de test. En outre, l'écart de performance entre les différentes configurations du c-vecteur est beaucoup moins faible. la précision minimale atteinte sur les données de test est de 78,9 % contre 60,9 % avec les espaces AT (voir figure 3-(b)).

taille du c-vecteur	*DEV*			*TEST*		
	Nb gaussiennes dans le GMM-UBM					
	32	*64*	*128*	*32*	*64*	*128*
80	80,6	82,3	83,1	79,2	81,0	80,4
100	81,7	**84,6**	83,1	78,9	**82,3**	80,4
120	84,0	81,7	82,3	80,4	79,2	81,8

TABLE 1 – Précision de classification en thèmes (%) dans l'espace de variabilité totale avec différentes tailles de l'UBM et des c-vecteurs

4.2 Impact de la compression LTS

Les résultats obtenus en utilisant la représentation LTS sont présentés dans la figure 4. Afin de comparer les performances des différentes approches (AT/c-vecteurs/LTS), les meilleurs résultats sont présentés dans le tableau 2. Il faut préciser que ces résultats sont obtenus dans des conditions d'application "réelles", la configuration optimale (nombre de thèmes dans l'espace thématique) étant choisie dans la phase de développement. En conséquence, il pourrait exister un meilleur point de fonctionnement pour le corpus de test, ce qui explique l'écart entre les résultats répertoriés dans le tableau 2 et ceux des figures 3 et 4. Nous remarquons que la représentation LTS surpasse chacune des représentations AT (baseline) et c-vecteur sur le corpus de développement (figure 4-(a)) et de test

(figure 4-(b)) avec respectivement une précision de 89,7 % (+2, 7 points) et 85,3 % (+4, 6 points). En outre, les courbes du modèle LTS sont également plus stables et plus robustes que celles de la représentation c-vecteur. En effet, l'écart entre les deux valeurs extrêmes, qui est de 3,3 points avec les c-vecteurs, est égal à 2,1 points pour la représentation LTS sur les données de test.

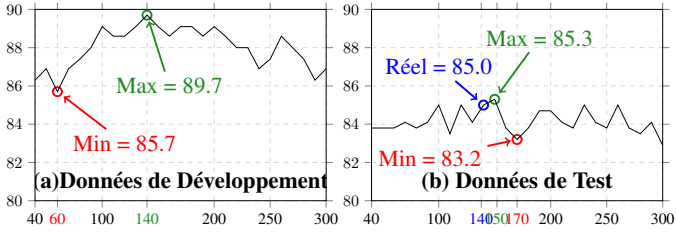


FIGURE 4 – Précisions de la classification en thèmes (%) en utilisant un vecteur compact de représentations LTS sur les corpus de développement et de test avec différentes configurations expérimentales. L'axe X représente le nombre des valeurs propres m ($40 \leq m \leq 300$).

Approche de représentation	**DEV**		**TEST**
	taille	**prec. %**	**prec. %**
Modèle AT (baseline)	202	86,3	80,4
Modèle AT + c-vecteur	100	84,6	82,3
Modèle ATAT + LTS	140	89,7	**85,0**

TABLE 2 – Précision de la classification en thèmes (%) sur le corpus de test en utilisant la configuration optimale obtenue sur le corpus de développement.

5 Conclusion

La performance des SRAP dépend des conditions d'enregistrement et la qualité des transcriptions automatiques produites influe sur les tâches de compréhension du langage. Ce papier propose une solution pour faire face aux erreurs de transcription en projetant un dialogue dans un sous-espace de caractéristiques robustes appelé Sous-espace Thématique Latent (LTS). Les expériences conduites dans le cadre de la classification de conversations a montré l'efficacité du modèle LTS proposé par rapport à l'utilisation des représentations c-vecteur et AT classiques. Cette représentation haut-niveau permet d'améliorer considérablement la performance de la tâche d'identification de thèmes avec un gain supérieur à 3 et 2 points respectivement par rapport à l'utilisation des représentations AT et c-vecteur. Le modèle LTS est combiné avec une compression PCA vu la faible taille des données (740 dialogues dans le corpus d'apprentissage). Dans la continuité de ce travail, il est intéressant d'évaluer cette représentation prometteuse avec des données plus volumineuses que celles du projet DECODA. Ainsi, d'autres méthodes de compression comme l'auto-encodeur et la PCA Probabiliste pourraient pbtenir des résultats meilleurs durant les différentes tâches d'analyse de la parole.

Références

ABDI H. & WILLIAMS L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews : Computational Statistics*, **2**(4), 433–459.

BECHET F., MAZA B., BIGOUROUX N., BAZILLON T., EL-BEZE M., DE MORI R. & ARBILLOT E. (2012). : LREC'12.

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, **3**, 993–1022.

BOUSQUET P.-M., MATROUF D. & BONASTRE J.-F. (2011). Intersession compensation and scoring methods in the i-vectors space for speaker recognition. In *Interspeech*, p. 485–488.

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.

EISENSTEIN J. & BARZILAY R. (2008). Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 334–343 : ACL.

GOLUB G. H. & REINSCH C. (1970). Singular value decomposition and least squares solutions. *Numerische mathematik*, **14**(5), 403–420.

HAZEN T. (2011). Topic identification. *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*, p. 319–356.

LAGUS K. & KUUSISTO J. (2002). Topic identification in natural language dialogues using neural networks. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, p. 95–102, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics.

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

MAY C., FERRARO F., MCCREE A., WINTRODE J., GARCIA-ROMERO D. & VAN DURME B. (2015). Topic identification and discovery on text and speech.

MELAMED I. & GILBERT M. (2011). Speech analytics. *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*, p. 397–416.

MORCHID M., BOUALLEGUE M., DUFOUR R., LINARÈS G., MATROUF D. & DE MORI R. (2014a). I-vector based approach to compact multi-granularity topic spaces representation of textual documents. In *EMNLP* : SIGDAT.

MORCHID M., BOUALLEGUE M., DUFOUR R., LINARÈS G., MATROUF D. & DE MORI R. (2014b). I-vector based representation of highly imperfect automatic transcriptions. In *Conference of the International Speech Communication Association (Interspeech) 2014* : ISCA.

MORCHID M., DUFOUR R., BOUALLEGUE M. & LINARÈS G. (2014c). Author-topic based representation of call-center conversations. In *SLT* : IEEE.

PURVER M. (2011). Topic segmentation. *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*, p. 291–317.

ROSEN-ZVI M., GRIFFITHS T., STEYVERS M. & SMYTH P. (2004). The author-topic model for authors and documents. In *Uncertainty in artificial intelligence*, p. 487–494 : AUAI Press.

TUR G. & DE MORI R. (2011). *Spoken language understanding : Systems for extracting semantic information from speech*. John Wiley & Sons.

XING E. P., JORDAN M. I., RUSSELL S. & NG A. (2002). Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*.

Stratégies d'adaptation de la vitesse d'articulation lors de conversations spontanées entre locuteurs natifs et non-natifs

Barbara Kühnert[1] Tanja Kocjančič Antolík[1]
(1) Laboratoire de Phonétique et Phonologie (UMR 7018, CNRS-Sorbonne Nouvelle)
19 rue des Bernadins, 75005 Paris, France
barbara.kuhnert@univ-paris3.fr, tanja.kocjancic@univ-paris3.fr

RESUME

Cet article examine la vitesse d'articulation dans un corpus de conversations spontanées entre locuteurs natifs et non-natifs. L'objectif est d'étudier (i) dans quelle mesure les locuteurs natifs adaptent dans leur L1 leur vitesse d'articulation aux apprenants L2 et (ii) dans quelle mesure les deux locuteurs en interaction ont tendance à rapprocher ou à dissocier leurs caractéristiques temporelles au cours d'une conversation. Les données proviennent du corpus SITAF d'interactions tandem en anglais-français. A ce jour, 10 sujets ont été analysés, chacun ayant été enregistré dans trois conditions différentes : en utilisant sa L1 avec un autre locuteur natif, en utilisant sa L1 avec un apprenant L2, et en utilisant sa L2 avec un interlocuteur parlant sa propre L1. Les résultats indiquent que les propriétés rythmiques de la L1 ont une nette influence sur les variations de la vitesse d'articulation des locuteurs non seulement lorsqu'ils interagissent dans leur L2 mais également dans leurs stratégies d'adaptation lorsqu'ils interagissent avec des apprenants.

ABSTRACT

Adaptation of articulation rate in spontaneous speech between native speakers and L2 learners.

The present study explores the patterns of articulation rate (AR) in spontaneous conversations between native speakers and L2 learners. More specifically, we were interested in (i) whether and how native speakers accommodate their L1 AR in conversation with L2 learners, and (ii) whether and how the participants show convergent or divergent patterns of AR in the course of a conversation. The data were taken from the SITAF corpus of English/French tandem interactions. To date 10 subjects have been analyzed, each recorded in three different conditions: when using his/her L1 with another native speaker, when using his/her L1 with an L2 learner, and when using his/her L2. The results suggest that both, rate adjustments to foreign learners as well as L2 rate variations, are strongly influenced by the language-specific properties of the L1 background.

MOTS-CLES : acquisition, langue seconde, production, vitesse d'articulation, adaptation phonétique, français, anglais américain

KEYWORDS: second language acquisition, speech production, articulation rate, phonetic accommodation, French, General American

1 Introduction

Dans cet article nous étudions la vitesse d'articulation lors de conversations spontanées en tandem entre des locuteurs du français et de l'anglais-américain. La vitesse d'articulation est une des variables prosodiques de l'organisation temporelle de la parole. Elle est supposée représenter le rythme avec lequel les segments sont produits. Elle est généralement définie comme une mesure rythmique qui calcule le nombre de segments réalisés par seconde, les pauses ayant été exclues. Cette exclusion des pauses évite de prendre en considération des caractéristiques individuelles propres aux locuteurs comme les pauses silencieuses, les hésitations ou d'autres réalisations affectives.

Dans le domaine de l'acquisition d'une langue seconde ou étrangère (dorénavant L2), la vitesse d'articulation (VA) est souvent associée aux notions de compétence et d'intelligibilité (Trouvain et Möbius, 2014). Une progression dans la maîtrise d'une L2 serait reflétée par une VA plus constante et plus rapide. Ainsi, les apprenants d'une L2 ont été décrits comme ayant une VA plus lente dans leur L2 que dans leur L1 (Gut, 2009). Néanmoins, la littérature fournit des résultats plus mitigés concernant les différences de VA entre les locuteurs natifs et non-natifs. D'un côté, il a été observé que les locuteurs L2 ont une VA clairement plus lente que les locuteurs natifs (Gut, 2009, pour les apprenants anglophones de l'allemand). De l'autre, il a été rapporté que celle des locuteurs natifs et non-natifs ne présente pas de différences significatives (Schwab et al., 2012 ; Avanzi et al., 2012, pour les apprenants suisse-allemands du français).

Un autre concept pertinent dans le contexte de l'acquisition d'une L2 est la notion d'adaptation – aussi appelée accommodation, alignement ou entraînement. Dans le sens le plus large du terme, l'adaptation phonétique désigne les modifications des caractéristiques vocales en réponse à une situation communicative ou à un interlocuteur. Les études initiales ont considéré ce phénomène soit comme un processus automatique soit comme une stratégie d'interaction volontaire et sociale. Les recherches plus récentes suggèrent cependant qu'un couplage entre les différentes approches existantes est plus approprié pour mieux comprendre ce phénomène (voir Babel, 2009).

La parole de locuteurs natifs adressée à des apprenants, appelée également 'foreigner-directed speech' est souvent caractérisée comme ayant « un rythme plus lent » (Ellis, 1997:45). Toutefois, peu d'auteurs ont étudié les propriétés acoustiques d'une parole utilisée dans des interactions entre locuteurs natifs et non-natifs. Bien que Biersack et al. (2005) aient montré un ralentissement dans le débit de la parole destinée à des interlocuteurs L2, cela résultant principalement d'un allongement de la durée des pauses. Quant à Uther et al. (2007), ils n'ont trouvé aucun corrélat temporel dans leur étude. En comparant la parole adressée à un interlocuteur imaginaire et à un interlocuteur réel, Scarborough et al. (2007) ont observé que les participants ont adapté leur VA dans les deux cas, mais que cette adaptation était plus marquée dans la situation avec interlocuteur imaginaire.

D'un point de vue plus dynamique, les études portant sur la convergence des réalisations phonétiques s'intéressent plutôt aux adaptions entre deux individus au cours d'une interaction (voir Pardo, 2013). Pour ce qui est de la VA, les études menées aboutissent à des résultats variables et contradictoires. Street (1982), par exemple, a observé une convergence dans la VA entre des locuteurs natifs qui participaient à un entretien. En revanche lors d'une étude plus récente dans laquelle les sujets étaient impliqués dans une 'map task', aucune convergence rythmique n'a été observée (Pardo et al., 2010). Concernant les interactions mixtes entre locuteurs natifs et non-natifs, les études ont montré que les locuteurs ont tendance soit à se rapprocher, soit à accentuer leurs différences. En se basant sur des jugements perceptifs, Kim et al. (2011) affirment que la convergence phonétique entre deux interlocuteurs varie en fonction de la distance entre les langues en présence. Lorsque la distance entre

les langues est réduite (comme par exemple entre des locuteurs natifs de l'anglais américain), les locuteurs sont plus susceptibles de converger que lorsque leurs langues sont plus éloignées, c'est-à-dire entre locuteurs de deux dialectes différents ou de deux langues différentes (des locuteurs natifs de l'anglais américain s'adressant à des apprenants coréens de l'anglais). Toutefois, cette hypothèse n'a été que partiellement confirmée dans une étude acoustique subséquente (Rao, 2013) ; des sujets natifs de l'anglais américain ont effectivement montré une adaptation réciproque alors que les participants inter-dialectaux (américain et anglais indien) et inter-langues (anglais américain et apprenants hispanophones) ont montré à la fois de la convergence et de la divergence.

Les interactions en tandem entre des paires de locuteurs de L1 différente peuvent s'avérer fort utiles pour tenter d'éclaircir certaines des conclusions contradictoires rapportées dans la littérature. L'apprentissage des langues en tandem est basé sur le principe de réciprocité selon lequel des paires de locuteurs de langues différentes visent à apprendre la langue de l'autre. Les deux partenaires contribuent de manière égale en termes de réalisations L1 / L2. Ainsi, les interactions en tandem permettent l'évaluation de la VA des *mêmes* locuteurs, parlant des *mêmes* sujets de conversation, dans le *même* cadre expérimental, mais dans deux langues différentes. Notre travail visait à répondre aux questions suivantes: (i) Comment les locuteurs natifs adaptent-ils leur VA dans leur L1 lors de conversations avec des apprenants parlant leur langue ; (ii) les partenaires tandem montrent-ils des stratégies d'adaptation de la VA au cours d'un échange ; et (iii) comment la VA en L1 se compare-t-elle à celle en L2 pour un locuteur donné ?

2 Méthode

2.1 Participants et collecte de données

Pour l'étude présentée ici, nous utilisons une partie du corpus SITAF (Horgues et Scheuer, 2014), un corpus d'enregistrements audio/vidéo d'interactions en tandem anglais-français. A ce jour, nous avons analysé cinq pairs de tandems, chacune constituée d'un(e) étudiant(e) francophone en Licence à l'Université Paris 3-Sorbonne Nouvelle et d'un(e) étudiant(e) anglais-américain(e) en échange dans la même université. L'âge des étudiants francophones varie entre 17 et 21 ans, celui des étudiants américains entre 19 et 20 ans. Les conversations ont été enregistrées numériquement dans une chambre insonorisée à l'Université de Paris 3. Chaque locuteur a participé à trois types différents d'interaction conversationnelle :

- une conversation dans sa L1 avec un autre locuteur de même L1 (interaction contrôle L1),
- une conversation avec le partenaire tandem dans laquelle le locuteur cible utilise sa L1, mais son interlocuteur parle dans sa L2 (interaction tandem L1),
- une conversation avec le même partenaire tandem dans laquelle le locuteur cible utilise sa L2 (interaction tandem L2).

Avec cette configuration, chaque francophone a été étudié lorsqu'il parle en français avec un autre francophone (interaction contrôle L1), lorsqu'il parle en français avec un anglophone ayant le français comme L2 (interaction tandem L1), et lorsqu'il parle en anglais (donc dans sa L2) avec un anglophone dans sa L1 (interaction tandem L2). La même répartition existe pour les anglophones L1. Chaque conversation, à son tour, était composée de deux activités communicatives différentes : une activité de narration et une activité d'argumentation, cela permettant d'assurer la réciprocité du dialogue. La durée moyenne des conversations était de 7:08 min et, au total, nous avons examiné 4h00 d'interactions environ.

2.2. Méthode d'analyse

Les conversations ont été transcrites orthographiquement à l'aide de Transcriber. Celles en anglais ont ensuite été alignées avec le système MAuS (Schiel et al., 2011), puis vérifiées et corrigées manuellement dans PRAAT (Boersma et Weenink, 2014). Quant à celles en français, elles ont été annotées manuellement par les auteurs à l'aide de PRAAT. Ont été exclus des différentes conversations pour l'analyse les passages contenant des bruits de fond, des rires, des chevauchements ou des disfluences, pour ne conserver que des unités produites de façon régulière ('fluent stretches'). On définit ici une unité comme un intervalle de 3 à 20 syllabes entre deux pauses. Une pause correspond à un intervalle silencieux de plus de 200 ms. Les unités de moins de 3 syllabes ont été éliminées car elles contenaient principalement des répliques courtes, telles que *yes*, *you know* ou *oui*, *c'est vrai* ; les unités de plus de 20 syllabes ont été éliminées car elles étaient peu nombreuses et inégalement réparties. Au total 1475 unités ont été prises en compte dans l'analyse finale. La VA a été calculée en divisant la durée de chaque unité par le nombre de syllabes effectivement produites, c'est-à-dire le nombre de syllabes phonétiquement réalisées. La figure 1 présente les données brutes de VA pour tous les locuteurs dans les conversations tandem qui seront analysées dans ce travail. Il suffit ici de noter que la majorité des unités compte 12 syllabes ou moins et que les unités courtes n'étaient pas produites uniquement par les apprenants L2.

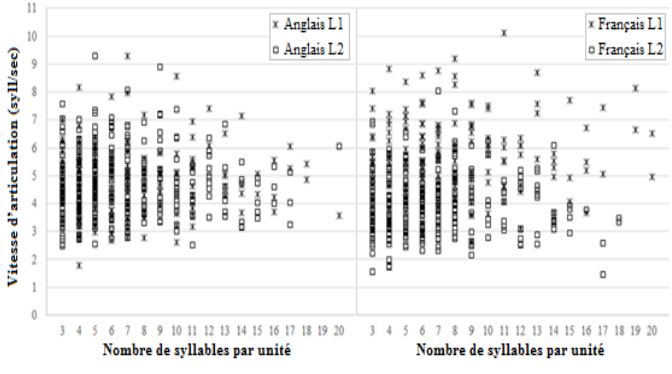


FIGURE 1 : Données brutes de la vitesse d'articulation (syll/sec) en fonction du nombre de syllabes par unité. A gauche : interactions en tandem en anglais ; à droite : interactions en tandem en français.

L'analyse statistique des données a été réalisée en appliquant des modèles linéaires à effets mixtes (lme4 package, Douglas et al., 2015) dans R (R Core Team, 2012) pour trois sous-ensembles de données : (i) les interactions contrôle en L1, (ii) les interactions avec le locuteur cible dans sa L1 (tandem L1), et (iii) les interactions où le locuteur cible parle dans sa L2 avec le partenaire du tandem (tandem L2). La vitesse d'articulation constitue la variable dépendante et a été testée séparément avec les prédicteurs suivants pour les trois sous-ensembles : (i) *Langue L1* (anglais/français) ; (ii) *Langue L1* (anglais/français), *Séance* dans laquelle la L1 a été utilisé (contrôle/tandem L1), et l'interaction des facteurs *Langue L1* / *Séance* ; et (iii) *Langue* (anglais/français), le *Statut* de la langue (parlée comme L1 ou L2), et l'interaction des facteurs *Langue* / *Statut.*

3 Résultats

3.1 Stratégies globales d'adaptation

La figure 2 présente la VA en fonction de chaque langue et de chaque type d'interaction. On note d'abord une différence importante de VA entre les locuteurs francophones et anglophones dans les conversations contrôle ($\chi2(1)=12.855$, $p<0,001$). Avec une moyenne de 6,78 syll/sec, les sujets français parlent significativement plus vite que les sujets américains (moyenne de 4,88 syll/sec).

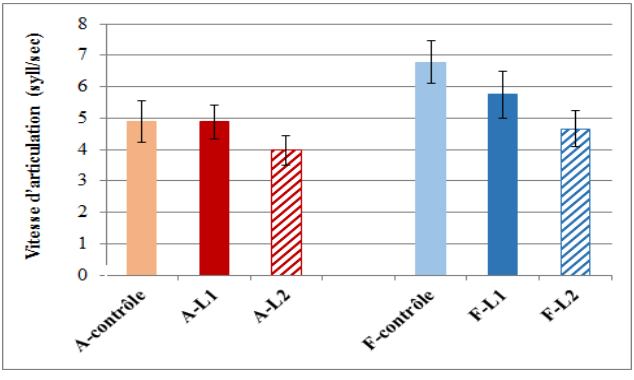


FIGURE 2 : Moyenne et écart type de la VA (syll/sec), de gauche à droite : Sujets américains en interaction contrôle (A-contrôle), en interaction tandem L1, c'est-à-dire parlant anglais (A-L1) et en interaction tandem L2, c'est-à-dire parlant français (A-L2) ; suivi par les sujets français en interaction contrôle (F-contrôle), en interaction tandem L1 (F-L1), et en interaction tandem L2, c'est-à-dire parlant anglais (F-L2).

De même, on observe un effet significatif de *Langue L1* ($\chi2(1)=9.439$, $p<0,01$), de *Séance* ($\chi2(1)=25.201$, $p<0,001$), et, surtout, une interaction significative entre *Langue L1* et *Séance* ($\chi2(1)=23.718$, $p<0,001$) lors d'une comparaison de l'utilisation de la L1 dans les interactions contrôle et dans les interactions en tandem L1. Seuls les sujets francophones montrent une adaptation importante quand ils s'adressent aux locuteurs non-natifs et ralentissent leur vitesse d'articulation de 6,78 syll/sec à 5,75 syll/sec. Aucune adaptation n'a pu être observée de la part des locuteurs anglais-américain. De fait, la VA utilisée par les anglophones en anglais dans les deux conditions (contrôle et avec des non-natifs) est identique (4,88 syll/sec).

Finalement, nous avons comparé la VA de la L1 avec celle de la L2 lors des sessions tandem (tandem L1 et tandem L2). Cette analyse n'a pas révélé d'effet de *Langue* ($\chi2(1) = 0,373$, $p = 0,54$), mais un effet significatif du *Statut* de langue (parlée comme L1 ou L2) ($\chi2(1)=169.09$ $p<0,001$), ainsi qu'une interaction significative entre *Langue* et *Statut* ($\chi2(1)=6.388$, $p<0,05$). Les locuteurs des deux groupes sont plus rapides dans leur langue maternelle que dans leur L2. Cependant, dans les tandems en langue anglaise les apprenants français n'étaient pas significativement plus lents que les natifs américains, comme le montre la figure 2 (F-L2 : 4,65 syll/sec versus A-L1 : 4,88 syll/sec). Dans les tandems en langue française, par contre, les apprenants anglophones montrent des valeurs plus basses que les partenaires français (A-L2 : 3,97 syll/sec versus F-L1 : 5,75 syll/sec) - même si, comme indiqué ci-dessus, ces derniers avaient déjà ralenti considérablement leur VA afin de s'adapter à leurs interlocuteurs.

3.2 Stratégies dynamiques d'adaptation

Pour étudier le phénomène de convergence phonétique et examiner dans quelle mesure les locuteurs tentent de rapprocher leurs caractéristiques temporelles à celles de leur interlocuteur au cours d'une interaction, nous avons suivi une méthode utilisée par Levitan et Hirschberg (2011). Comme indiqué précédemment, chaque session tandem comprenait deux activités conversationnelles, chacune correspondant en gros à la moitié d'une interaction. Nous avons alors calculé la moyenne de la VA pour les deux activités séparément et comparé les différences de VA entre la première et deuxième partie pour les binômes individuels. Nous supposons qu'il y a eu convergence lorsque la similitude entre les deux participants était plus importante (i.e. la différence plus petite) au cours de la deuxième qu'au cours de la première activité.

La table 1 regroupe les résultats obtenus. On note une tendance générale à la convergence pour tous les participants même si le degré de convergence dépend des paires étudiées. Dans les séances tandem en langue française, cette tendance est presque exclusivement due à une réduction de la VA de la part des locuteurs natifs francophones. Ainsi l'adaptation des participants français abordée précédemment, c'est-à-dire la réduction de leur VA pour s'aligner sur celle de leurs interlocuteurs non-natifs, semble refléter un ralentissement progressif au cours d'une interaction. Pendant les sessions tandem en langue anglaise, les patrons de convergence temporelle sont plus hétérogènes. Les sujets natifs américains ainsi que les non-natifs français ont montré des ajustements mutuels.

Les résultats démontrent que les sujets adaptent leur VA mais que cette adaptation dépend de l'interaction et de la langue qu'ils utilisent. Par exemple, en regardant la paire F5/A5 dans le tableau 1, on observe que le sujet F5 présente une diminution importante de sa VA pendant le tandem en français, avec très peu de changement de la part du partenaire A5. En revanche, pendant le tandem en anglais, c'est le sujet F5 qui montre peu ou pas de changement alors que c'est le sujet A5 qui augmente sa vitesse d'articulation pour s'aligner vers celle de son interlocuteur.

Paires	**Conversation**					
	Tandem anglais			*Tandem français*		
	Jeu 1	Jeu 2	Différence	Jeu 1	Jeu 2	Différence
F1 A1	4,65 5,20	4,97 4,16	**0,55 > - 0,23**	5,62 3,92	5,36 3,82	**1,70 > 1,54**
F2 A2	4,07 5,06	4,39 5,10	**0,99 > 0,71**	5,28 3,84	5,04 3,78	**1,44 > 1,26**
F3 A3	4,53 6,19	4,73 5,50	**1,66 > 0,77**	5,30 4,78	4,25 4,70	**0,52 < - 0,35**
F4 A4	4,85 4,54	4,81 4,57	**0,31 > 0,24**	7,47 3,82	6,64 4,02	**3,65 > 2,62**
F5 A5	4,88 4,35	4,85 4,72	**0,53 > 0,13**	5,20 3,58	4,80 3,48	**1,62 > 1,32**

TABLE 1 : Moyennes et différences (syll/sec) de VA pour chaque binôme. Dans la colonne « Différence » le premier chiffre donne la différence de la 1ère activité, le deuxième chiffre la différence de la 2ème activité.

4 Discussion

Même si une comparaison directe entre les deux langues n'était pas au centre de l'étude, nos résultats confirment ceux obtenus par des études antérieures sur les VA de langues différentes (Pellegrino et al., 2011). Les locuteurs francophones ont une VA plus rapide que les locuteurs américains. Cette différence peut être partiellement attribuée à des facteurs phonologiques propres à la langue. De fait, il est bien établi que la VA est influencée par de nombreux facteurs, tels que la complexité des syllabes, la structure phonotactique ou les phénomènes de réduction des syllabes non-accentuées (voir Schiering 2007). Comme le français présente des structures syllabiques plus simples et des groupes consonantiques moins complexes, l'articulation des syllabes peut se faire plus rapidement.

Nos résultats confirment également le fait que les sujets sont généralement plus lents dans leur L2 que dans leur L1 (cf. Gut, 2009). Cependant, nos données ne permettent pas de soutenir que la VA des locuteurs non-natifs est nécessairement plus lente que celle des locuteurs natifs. Bien qu'il en soit ainsi pour les apprenants américains – leur VA étant significativement plus lente que celle des français natifs – aucune différence au niveau du groupe n'a été notée entre les sujets francophones et américains dans les sessions tandem en langue anglaise. En effet, trois des cinq sujets français ont montré des VA plus élevées que les sujets natifs. Cette observation suggère que les variations dans la VA en L2 ne sont pas seulement liées à la VA en L1 sur le plan individuel (Trouvain et Möbius, 2014), mais dépendent également de la langue en question. La VA intrinsèquement plus élevée du français, qui découle des caractéristiques phonologiques mentionnées précédemment, se combine avec les différents taux individuels. Ceci permet aux apprenants francophones de s'approcher étroitement du niveau des locuteurs natifs alors que ce n'est pas le cas dans la situation inverse (à savoir les anglophones parlant français avec des natifs). Un effet similaire a été rapporté par Kim et al. (2013) dans une étude récente sur les variations de la VA parmi des locuteurs bilingues de langues diverses. Les auteurs ont trouvé des différences de VA entre les langues (avec, par exemple, le turc et le mandarin plus rapides que le coréen ou l'espagnol) et ces différences entre groupes linguistiques se sont également retrouvées dans les variations de VA en anglais L2.

En outre, nous avons observé que les stratégies d'adaptation entre locuteurs natifs et non-natifs semblent être influencées par la langue. Au niveau global, les locuteurs français ont ralenti leurs productions de manière significative quand ils s'adressaient à des partenaires non-natifs. Aucune adaptation de ce genre n'a été retrouvée de la part des locuteurs américains lors des sessions en anglais. Comme nous l'avons remarqué, peu de travaux se sont concentrés sur les propriétés acoustiques du 'foreigner talk' et, à notre connaissance, aucune étude n'a comparé les paroles adressées aux apprenants dans différentes langues. Par conséquent, nous ne pouvons qu'émettre l'hypothèse que ces ajustements unilatéraux sont, à nouveau, guidés par la VA sous-jacente plus rapide du français, favorisant une réduction du rythme conversationnel.

Quant à la convergence phonétique au sein d'une même session, nous avons constaté pour la majorité des paires une plus grande similitude de la VA dans la deuxième partie des interactions. Il est intéressant de noter que cette convergence ne révèle pas une dynamique de type 'L1 leader / L2 suiveur' comme nous aurions pu le croire. Trouvain et Moebius (2014), par exemple, ont constaté que les apprenants ont augmenté leur VA en L2 lors de la lecture de phrases après avoir été exposés directement à une phrase modèle lu par un locuteur natif. Dans les interactions spontanées de la présente étude, ce sont plutôt les locuteurs natifs qui se sont adaptés lors des conversations en français, alors qu'en anglais les deux participants convergent, indépendamment du statut de la langue parlée. En d'autres termes, l'existence d'une modification de la VA d'un locuteur est déterminée par la langue utilisée et le rôle pris dans l'interaction tandem. Ceci étant, on ne peut pas formuler d'hypothèses sur

des raisons purement automatiques qui déclencheraient la convergence (cf. Pardo, 2013). Une analyse plus détaillée de la convergence pendant les interactions tandem sera nécessaire à l'avenir.

Un facteur que nous n'avons pas inclus dans notre analyse est la longueur des unités ('phrase length'). Or, Quené (2008) a proposé que la VA varie en fonction du nombre de syllabes par unité et faisant, par conséquent, de la longueur des unités un facteur déterminant pour le rythme de la parole. Les unités plus courtes seraient produites plus lentement que les unités plus longues en raison du phénomène de 'raccourcissement anticipé' ('anticipatory shortening'). Ainsi, les locuteurs appliquent une plus grande vitesse et raccourcissent de fait leurs syllabes s'ils anticipent qu'elles seront en plus grand nombre dans l'unité produite. Cependant, cette tendance n'a pas pu être confirmée dans une étude sur des discours spontanés en allemand (Trouvain et al., 2001) ni dans une étude sur la parole spontanée dans deux variétés distinctes de l'anglais-américain (Jacewicz et al., 2010). Comme le montre la figure 1, les sujets ont réalisé une grande variété d'unités courtes et longues en utilisant leur L1 et leur L2, avec une légère tendance pour les locuteurs natifs de réaliser plus d'unités longues que les non-natifs. Néanmoins, il ne semble pas y avoir de corrélation stricte entre VA et longueur des unités. En fait, l'analyse a montré que la VA des sujets francophones correspondait étroitement à la performance des locuteurs natifs, *malgré* le fait qu'ils aient réalisé plus d'unités courtes dans les interactions en anglais. Aussi, la non-prise en compte de la longueur des unités dans notre analyse ne devrait pas avoir affecté l'interprétation de nos résultats.

En résumé, à partir de cette étude, plusieurs conclusions peuvent être formulées sur la VA lors de conversations spontanées entre locuteurs français et américains. Comme observé dans des études précédentes, la VA est nettement plus rapide en français qu'en anglais-américain. Tous les sujets sont plus lents dans leur L2. Une comparaison entre la VA et les stratégies d'adaptation entre sujets natifs et non-natifs suggère une influence majeure des caractéristiques temporelles inhérentes aux langues. Seuls les sujets français ont parlé en L2 avec un VA à peu près égale à celle des natifs américains, et eux seuls aussi ont eu tendance à clairement adapter leur VA à celle de leurs interlocuteurs étrangers. Ainsi, les patterns de convergence phonétique de la VA varient selon les paires étudiées et la langue utilisée.

Remerciements

Le travail présenté ici a été soutenu par le Labex "*Empirical Foundations in Linguistics*" (ANR-10-LABX-0083).

Références

AVANZI, M., DUBOSSON, P., SCHWAB, S. (2012). Effects of dialectal origin on articulation rate in French. *Proceedings of Interspeech 2012,* 651-654.

BABEL, M. E. (2009). *Phonetic and Social Selectivity in Speech Accommodation*. Doctoral Dissertation. Berkeley: University of California.

BIERSACK, S., KEMPE, V., KNAPTON, L. (2005). Fine Tuning speech registers: a comparison of the prosodic features of child-directed and foreigner-directed speech. Proceedings of the 9th European Conference on Speech Communication and Technology, 2401-2404.

GUT, U. (2009). *Non-native Prosody. A corpus-based analysis of the phonetic and phonological properties of L2 English and L2 German*. Frankfurt: Peter Lang.

HORGUES, C., SCHEUER, S. (2014). Why some things are better done in tandem? *Proceedings of the Third International Conference on English Pronunciation: Issues and Practices*, 41-44.

JACEWICZ, E., Fox, R., O'Neill, C., Salmons, J. (2009). Articulation rate across dialect, age, and gender. *Language Variation and Change*, 21, 233–256.

KIM, M., ACKERMAN, L., BURCHFIELD, L., DAWDY-HESTERBERG, L., LUQUE, J., MOK, K., BRADLOW, A. (2013). Rate variation as a talker-specific/language-general property in bilingual speakers. *The Journal of the Acoustical Society of America*, 133, 3574.

KIM, M., HORTON, W., BRADLOW, A. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *J. ournal of Laboratory Phonology*, 2, 125–156.

LEVITAN, R., HIRSCHBERG, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. *Proceedings of Interspeech 2011*, 3081-3084.

PELLEGRINO, F., COUPE, C., MARSICO, E. (2011). A cross-language perspective on speech

PARDO, J. S. (2013). Measuring phonetic convergence in speech production. *Frontiers in Psychology, 4*, 559.

PARDO, J. S., CAJORI JAY, I., KRAUSS, R. M. (2010). Conversational role influences speech imitation. *Attention, Perception, & Psychophysics*, 72, 2254-2264.

QUENÉ, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *Journal of the Acoustical Society of America*, 123, 1104-1113.

RAO, G. (2013). *Measuring phonetic convergence: segmental and suprasegmental speechadaptations during native and non-native talker interactions*. UT Electronic Theses and

SCARBOROUGH, R., BRENIER, J., ZHAO, Y., HALL-LEW, L., DMITRIEVA, O. (2007). An Acoustic Study of Real and Imagined Foreigner-Directed Speech. *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS XVI)*, 2165-2168.

TROUVAIN, J., MÖBIUS, B. (2014). Sources of variation of articulation rate in native and non-native speech: comparisons of French and German. *Proceedings of Speech Prosody 2014*, 275-279.

TROUVAIN, J., KOREMAN, J., ERRIQUEZ, A., BRAUN, B. (2001). Articulation rate measures and their relations to phone classification of spontaneous and read German speech. *Proceedings of the ISCA Workshop on Adaptation Methods for Speech Recognition*, 155-158.

SCHIERING, R. (2007). The phonological basis of linguistic rhythm. Cross-linguistic data and diachronic interpretation. *Sprachtypologie und Universalienforschung*, 60, 337–359.

SCHWAB, S., DUBOSSON, P., AVANZI, M. (2012). Etude de l'influence de la variété dialectale sur la vitesse d'articulation en français. *Actes des XIXè Journées d'Etudes de la Parole*, 521-528.

UTHER, M., KNOLL, M.A., BURNHAM, D. (2007). Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech. *Speech Communication*, 49, 2-7.

Stress, charge cognitive et signal de parole : étude exploratoire auprès de pilotes de chasse.

L. Stavaux[1], M. Albart[1], V. Delvaux[1,2], K. Huet[1], M. Piccaluga[1] et B. Harmegnies[1]
(1) Institut de Recherche en Sciences et Technologies du Langage, Service de Métrologie et Sciences du Langage, Université de Mons, Belgique
(2) Fond National de la Recherche Scientifique, Belgique
luc.stavaux@umons.ac.be

RÉSUMÉ

Cet article traite des effets de la charge cognitive sur la fréquence fondamentale de pilotes de F-16 placés dans un scénario de vol de nuit. La charge cognitive a été estimée à l'aide de paramètres liés à la tâche (hétéro-évaluation), à l'individu (anxiété, auto-évaluation du stress ressenti) et à la situation (simulation contrôlée). Nos résultats montrent que l'écart mélodique est un bon candidat pour évaluer le niveau de la charge cognitive, même si la relation entre eux présente des profils individuels spécifiques. La création d'une typologie des situations de communication, l'adjonction d'autres indices acoustiques et le croisement avec des données physiologiques constituent les perspectives de cette étude.

ABSTRACT

Stress, cognitive load and speech signal : an exploratory study among fighter pilots.

This article discusses the effects of cognitive load on the pitch of F-16 pilots placed in a night flight scenario. Cognitive load was estimated using parameters related to the task (external evaluation), the individual (anxiety, perceived stress) and the situation (controlled simulation). Our results show that melodic distance is a good candidate to assess the level of cognitive load, even if the relation between them exhibits specific individual profiles. The prospects of this study are : the creation of a typology of communicative situations, the addition of other acoustic cues and the cross-validation with physiological data.

MOTS-CLÉS : charge cognitive, fréquence fondamentale, écart mélodique, stress, aéronautique.

KEYWORDS: cognitive load, pitch, melodic variation, stress, aeronautics.

1 Introduction

L'étude présentée ici a été réalisée dans le cadre du projet BIOVOC [1], dont l'objectif est d'étudier les effets, sur le sujet aux commandes d'un système complexe, du stress, de la fatigue et de la surcharge cognitive. Les variations de l'état du sujet y sont suscitées via la manipulation de variables

1. Action de Recherche Concertée "BIOVOC : impacts vocaux et métaboliques, sur le sujet aux commandes d'un système complexe, de variations situationnelles susceptibles d'agir sur son état.", collaboration entre les laboratoires d'Analyses Pharmaceutiques (B. Blankert), de Biologie humaine et Toxicologie (J.-M. Colet) et de Phonétique (B. Harmegnies) de l'Université de Mons

situationnelles (p.ex. en simulateur de vol), et objectivées de façon croisée par analyse de biomarqueurs (fréquence cardiaque, cortisol salivaire, signatures métaboliques, etc.) et analyse des productions de parole.

Le signal de parole est porteur de multiples informations extra-linguistiques, ayant trait notamment à des composantes stables de l'identité du locuteur (origine régionale, âge, genre, etc.) ainsi qu'à ses composantes évolutives (états physique, cognitif, émotionnel).

Le milieu aéronautique constitue un précieux contexte d'étude de ces phénomènes. En effet, les pilotes d'aéronef sont engagés dans une tâche éminemment complexe, impliquant de nombreux échanges verbaux d'information, et susceptible d'affecter largement leur état (fatigue physique, surcharge cognitive, stress, etc.). En retour, l'état du pilote est en partie constitutif de ces "facteurs humains" à même d'affecter lourdement le déroulement d'un vol, d'où l'intérêt de l'industrie aéronautique et des autorités de contrôle pour le développement d'instruments de détection, en particulier du stress et de la fatigue engendrés par une charge cognitive importante, à partir des signaux de parole enregistrés dans le cockpit (Cockpit Voice Recordings, CVR ci-dessous).

Le projet BIOVOC s'inscrit dans une longue tradition, au sein de notre laboratoire, de recherches centrées sur les relations entre variables psychologiques et signal de parole (Harmegnies & Landercy, 1992; Ruiz *et al.*, 1996; Piccaluga *et al.*, 2007). Ainsi, dans une étude pionnière menée sur de la parole francophone, Ruiz *et al.* (1996) ont comparé les effets du stress en situation réelle - en l'occurrence à partir d'enregistrements CVR recueillis lors d'un accident d'avion - avec les effets du "stress cognitif" (au sens défini par Harmegnies & Landercy, 1992) induit en laboratoire. Leurs analyses ont démontré que la fréquence fondamentale (F0) était globalement le paramètre le mieux à même de distinguer les différents niveaux de stress, indépendamment de la situation ou du locuteur. Les paramètres spectraux aboutissaient, quant à eux, à des résultats parfois intéressants mais globalement moins systématiques. L'hypothèse généralement admise est que la réaction psychophysiologique à un agent stressant implique une augmentation du rythme cardiaque, de la tension musculaire et de la pression sous-glottique, qui tous trois sont susceptibles de causer une élévation de la F0 (Alvear *et al.*, 2013).

Dans leur conclusion, Ruiz *et al.* (1996) pointaient un certain nombre d'éléments à prendre en compte pour de futurs travaux sur le sujet : (i) la nécessité de mieux définir, et donc opérationnaliser, le concept de "stress" ; (ii) l'intérêt d'objectiver la variable "stress" et ses composantes via leurs marqueurs physiologiques ; (iii) l'importance à accorder à la variation inter-individuelle (notamment dans la résistance au stress) (Ruiz *et al.*, 1996, pp.126-128). Une récente revue de la littérature consacrée aux "indices vocaux du stress" fait écho à la plupart de ces recommandations (Giddens *et al.*, 2013). Ainsi, il ressort de celle-ci que malgré la diversité des paramètres acoustiques étudiés (F0, intensité, VOT, jitter, shimmer, débit de parole, balance spectrale, rapport signal/bruit, etc.), c'est le plus systématiquement à une élévation de la F0 que le stress est associé (Mendoza & Carballo, 1998; Rothkrantz *et al.*, 2004; Huttunen *et al.*, 2011), même si celle-ci n'est pas universellement observée (Johannes *et al.*, 2000; Van Lierde *et al.*, 2009). Parmi les facteurs susceptibles de rendre compte de la variabilité des résultats obtenus, on pointera la grande variété des sous-types de stress (physique, cognitif, émotionnel ; en laboratoire vs. en situation réelle, etc.) induits dans les études considérées, ainsi que les caractéristiques individuelles des sujets étudiés (en termes d'expertise par rapport à la tâche, de réponse physiologique au stress, voire de personnalité).

Cet article traite des effets de la charge cognitive sur la fréquence fondamentale de pilotes de F-16 placés dans un scénario de vol de nuit. La charge cognitive peut être définie comme correspondant : « à l'intensité du traitement cognitif mis en œuvre par un individu lorsqu'il réalise une tâche donnée dans un contexte particulier» (Chanquoy *et al.*, 2007, p.248). De nombreux facteurs interviennent dans

l'intensité du traitement cognitif. Le premier est déterminé par le nombre d'informations impliquées dans la tâche et les relations qu'elles entretiennent entre elles : plus elles sont nombreuses, plus la charge augmente. L'expertise, c'est-à-dire les compétences et les connaissances antérieures dont dispose l'individu au moment de réaliser la tâche, influence la charge cognitive : plus le sujet est expert, plus il dispose de traitements automatisés, ce qui amoindrit sa charge cognitive. Les stratégies mises en œuvre par l'individu pour accomplir la tâche impactent la charge cognitive : une personne peut ignorer certaines informations ou optimiser certains traitements. La présence d'interférences dans le contexte d'exécution de la tâche joue également un rôle. La nature des traitements cognitifs à opérer est cruciale : des processus cognitifs ou sensoriels différents sont plus facilement exécutés simultanément, ce qui contribue à réduire la charge cognitive. L'effort que l'individu consent à produire pour mener la tâche affecte la charge cognitive maximale qu'il peut endurer : celle-ci dépend de sa motivation et de sa fatigue. Enfin, la pression temporelle intervient aussi : plus le temps disponible pour mener la tâche à son terme est faible, plus la charge cognitive augmente (Barrouillet *et al.*, 2004; Chanquoy *et al.*, 2007; Martin *et al.*, 2013).

La charge cognitive résulte donc de l'action combinée de trois entités majeures : la tâche, l'individu et l'environnement. Dans cette étude, une attention particulière a été portée au contrôle de ces trois types de facteur. A l'instar d'autres auteurs (Harmegnies & Landercy, 1992; Hansen *et al.*, 2000), nous considérons qu'un niveau de charge cognitive élevé peut constituer une pression d'ordre psychologique induisant une réponse psychophysiologique appelée stress.

2 Méthodologie

Nous avons eu l'opportunité d'observer 3 pilotes de F-16 lors d'un vol d'entraînement en simulateur. Le simulateur de vol permet de recueillir des données dans un environnement contrôlé, tant en termes d'environnement acoustique que de déterminants de la tâche. Ainsi, la tâche à réaliser est définie à l'avance par un scénario enchaînant diverses phases de vol. De ce point de vue, le simulateur de vol peut être considéré comme un excellent compromis entre les études réalisées en laboratoire et celles réalisées en situation réelle. Ensuite, le contexte militaire permet de recruter des participants au profil relativement similaire (voir table 1), notamment en termes de qualifications et de carrière professionnelle, voire de personnalité. Nous avons complété l'information recueillie sur chaque participant à l'aide de deux outils : (i) l'anxiété trait, c'est-à-dire structurelle et l'anxiété état, liée à la situation, ont été mesurée avec le STAI-Y (Spielberger *et al.*, 1993) : les trois pilotes présentent un niveau moyen d'anxiété trait ; (ii) l'auto-évaluation des stratégies de coping en situation stressante a été menée avec le CISS (Endler & Parker, 1990) : les sujets recourent à des stratégies similaires pour réduire leur niveau de stress : ils se concentrent sur la tâche pour en résoudre les problèmes (pilotes 2 et 3) ou essaient de modifier la situation (pilote 1).

La simulation consistait en une mission d'interception lors d'un vol de nuit par deux chasseurs F-16, le sujet incarnant le second appareil (ailier). Le décours temporel de cette simulation est présenté à la figure 1. Tout d'abord, les deux appareils se dirigent vers l'aéronef inconnu et lui demandent de s'authentifier à plusieurs reprises (*interception*). Suite à la non réaction de ce dernier, une phase de combat s'engage (*combat aérien*) : alors que le sujet doit aligner l'avion ennemi dans son viseur afin de l'abattre, il est pris pour cible par un missile sol-air et doit exécuter une manœuvre d'évitement. Quelques secondes plus tard, tous les systèmes électriques de l'avion se coupent (*panne électrique*), plusieurs alarmes se déclenchent et le sujet se retrouve dans le noir complet. Rapidement, il doit

	Sujet 1	Sujet 2	Sujet 3
Âge	27 ans	25 ans	27 ans
Formation initiale	cadre auxiliaire BAC	cadre auxiliaire BAC	cadre de carrière BAC +5
Niveau de qualification	formation basique au combat	formation au combat terminée aucune mission	formation basique au combat
Nombre d'heures de vol sur F-16	150	350	200
Grade	capitaine	sous-lieutenant	capitaine
Niveau d'anxiété STAI-Y Trait	moyen	moyen	moyen
Stratégie de coping dominante	évitement distraction	problème tâche	problème tâche

TABLE 1 – Profil des pilotes composant l'échantillon

activer le générateur de secours (EPU) qui n'a pas démarré automatiquement. Une fois le courant rétabli, il se lance dans un premier diagnostic, lequel est interrompu par une panne de son dispositif d'affichage tête haute (HUD) (*panne affichage tête haute*). Après plusieurs tentatives infructueuses de redémarrage du générateur principal, le pilote doit se résoudre à retourner d'urgence à la base (*retour à la base*). L'autre appareil étant également endommagé, le sujet devient leader de la formation et est chargé de le guider "dans son aile" jusque la piste d'atterrissage. En raison de la panne de l'affichage tête haute et de l'absence de visibilité, l'approche finale de la piste se réalise aux instruments avec le balisage ILS. Le train d'atterrissage ne se verrouille pas correctement lors de la première sortie et nécessite une seconde tentative. Enfin, les freins de roue étant endommagés, le sujet doit libérer le crochet d'appontage pour saisir un câble afin de freiner l'appareil lors du roulage sur la piste. A l'instar de Huttunen *et al.* (2011), nous avons découpé ce scénario en plusieurs phases en fonction de la charge cognitive induite par la tâche.

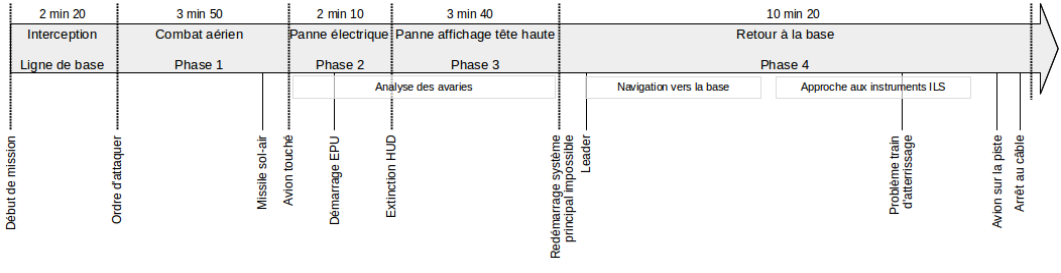


FIGURE 1 – Représentation chronologique du scénario de vol

Nous avons enregistré le signal de parole de chaque sujet grâce à un micro-casque de type Sennheiser HSP 2-EW-M relié à un dispositif d'enregistrement portable Sony Pro PCM-D100. L'ensemble des écrans d'informations et des instruments de vol ont été filmés depuis la salle de contrôle par un caméscope full HD Sony HDR 405. Pour des raisons de confidentialité, seuls les enregistrements sonores ont pu être analysés en dehors de l'enceinte militaire.

La charge cognitive effectivement induite par la tâche a été évaluée individuellement pour chaque

pilote, a posteriori, sur base du visionnage par un pilote instructeur de l'enregistrement vidéo de l'ensemble de l'exercice, phase de vol par phase de vol. Pour ce faire, nous avons employé le même outil que Huttunen *et al.* (2011) qui évalue trois dimensions. La première est appelée la conscience de la situation (CS) (Endsley, 1995) et comporte 3 niveaux : la perception des informations dans l'environnement ; la compréhension et l'intégration de celles-ci dans la situation actuelle et la projection de celles-ci afin d'anticiper l'évolution probable de la situation. La deuxième consiste en la charge informationnelle (CI) et comprend deux niveaux : la quantité et la qualité des informations à traiter. La troisième dimension s'axe quant à elle sur la charge décisionnelle (CD), 4 critères sont évalués : le temps disponible pour réaliser des choix ; la criticité de ceux-ci sur la poursuite de la tâche ; la quantité de décisions à prendre et la complexité des jugements à poser en fonction du nombre d'options disponibles. Le pilote instructeur a attribué, pour chaque sous-niveau, un score variant de 0 (aucune charge cognitive) à 100 (charge cognitive extrême) grâce à une échelle visuelle analogique. Cette hétéro-évaluation constitue notre mesure de la charge cognitive induite par la tâche.

Nous avons également demandé à chaque pilote d'estimer son niveau de stress ressenti pour chaque phase de vol, toujours à l'aide d'une échelle visuelle analogique variant de 0 (aucun stress) à 100 (beaucoup de stress). Cette auto-évaluation constitue notre mesure du stress provoqué par la charge cognitive. Par ailleurs, le niveau d'anxiété état des pilotes a été mesuré grâce au STAI-Y avant et après la simulation.

Nous avons extrait une mesure de la fréquence fondamentale toutes les 10 ms pour l'ensemble des segments voisés en recourant à l'algorithme d'extraction du pitch par auto-corrélation du logiciel Praat (version 6.0.05). Ces mesures ont fait l'objet d'une vérification manuelle sur base du spectrogramme par un expert afin de corriger toute donnée aberrante.

Chaque mesure fréquentielle a été convertie en valeur harmonique (*VH*) obtenue par le logarithme en base 2 du rapport entre la valeur fréquentielle exprimée en hertz (f) et une valeur de référence (r), la hauteur de la note de musique do_2, soit 131 Hz (Zwicker & Feldtkeller, 1981). Cette conversion (figure 2) offre deux avantages : elle exprime la relation harmonique entre deux valeurs, indépendamment de la fréquence de départ, et tend à normaliser la distribution des données.

$$\boxed{VH = \log_2\left(\frac{f}{r}\right)}$$

FIGURE 2 – Calcul de la valeur harmonique

Afin de rendre comparables les mesures de F0 recueillies, nous avons calculé un écart mélodique (*EM*) qui consiste en la différence entre chaque valeur harmonique et la valeur harmonique moyenne ($\overline{VH_{lb}}$) obtenue à partir de la ligne de base de chaque pilote (phase d'interception sur la figure 1).

3 Résultats

Tout d'abord, le niveau d'anxiété état a augmenté chez les trois pilotes, mais de façon différente. Partant d'un niveau faible, les sujets 1 et 2 ont vu leur anxiété augmenter [2] (de 37 à 42, de 37 à 48)

2. Les notes du STAI-Y sont comprises entre 20 et 80 (< 35 niveau d'anxiété très faible ; de 35 à 45 : niveau faible ; de 46 à 55 : niveau moyen ; de 56 à 65 : niveau élevé ; > 65 niveau très élevé).

alors que l'anxiété du troisième sujet, plus forte avant la simulation, s'élève peu (de 49 à 50). La hausse d'anxiété la plus importante concerne donc le pilote 2 (+11).

La figure 3 présente le niveau de stress ressenti par chacun des pilotes. Il croît tout au long du scénario mais à des phases différentes pour chacun d'eux : le pilote 1 déclare que son stress augmente au cours de la phase 4 alors que cette hausse est ressentie à la phase 3 pour le pilote 3. Le pilote 2 affirme avoir éprouvé une élévation de son stress lors des phases 2 et 4 ainsi qu'une légère baisse lors de la phase 3. Le stress perçu diffère donc d'un individu à l'autre mais est systématiquement plus élevé à la fin de la simulation.

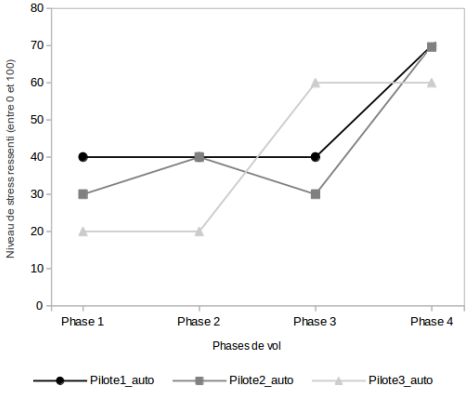


FIGURE 3 – Auto-évaluation du stress ressenti

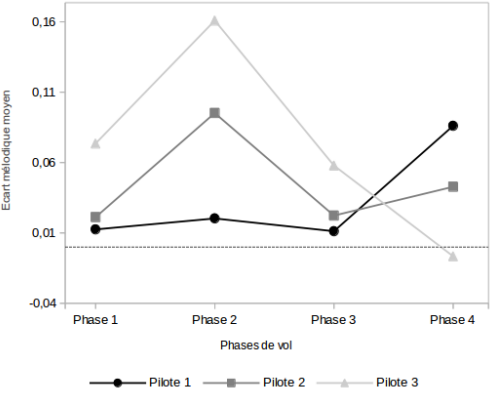


FIGURE 4 – Evolution des écarts mélodiques

En ce qui concerne l'hétéro-évaluation de la charge cognitive, les mesures positives de corrélation (r_{BP}) entre les 3 dimensions détaillées plus haut ($r_{CS\text{-}CI}$ = .76, $r_{CI\text{-}CD}$ = .69, $r_{CS\text{-}CD}$ = .49) nous ont conduits à les moyenner afin d'obtenir une mesure globale de la charge cognitive liée à la tâche pour chaque pilote. L'instructeur évalue la phase 3 comme la plus exigeante (> 75) et la phase 1 comme la moins astreignante (< 60) et cela quelque soit le pilote. Enfin, le niveau de stress ressenti par chaque pilote pour chaque phase est corrélé positivement à l'hétéro-évaluation de la charge cognitive (r_{BP} = .53).

La table 2 présente le nombre de valeurs de F0 extraites (toutes les 10 ms) et les valeurs moyennes des écarts mélodiques ($\overline{EM}$) pour chaque pilote et chaque phase, calculés par rapport à la valeur harmonique moyenne de leur ligne de base ($\overline{VH_{lb}}$). L'analyse de la figure 4 montre que, hormis pour la phase 4 chez le pilote 3, l'écart mélodique moyen est toujours supérieur à la ligne de base (trait discontinu sur la figure 4). Il augmente, dans des proportions différentes, chez tous les pilotes lors de la phase 2 et redescend lors de la phase 3 à un niveau proche de celui de la phase 1. Enfin, on observe des évolutions de l'écart mélodique opposées entre le pilote 3 et les pilotes 1 et 2 : la fréquence fondamentale du troisième sujet chute sous sa ligne de base tandis que celle des deux autres sujets grimpe et atteint son maximum chez le pilote 1 en fin de simulation.

Une analyse de variance à 2 critères croisés à modèle mixte (phase : facteur fixe ; pilote : facteur aléatoire) met en avant une absence de différences significatives entre les pilotes (F = .575, p = .591) et entre les phases (F = .484, p = .706) mais confirme la présence d'un effet d'interaction entre les phases et les pilotes (F = 258.311, p < .001). Les pilotes présentent donc des profils différents d'évolution de leur écart mélodique en fonction des phases, principalement en fin de simulation. Par

ailleurs, l'écart mélodique est faiblement corrélé à l'hétéro-évaluation de la charge cognitive ($r_{CC\text{-}EM}$ = .35), et à l'auto-évaluation du stress ($r_{CC\text{-}SC}$ = .17).

	Ligne de base	**Phase 1**	**Phase 2**	**Phase 3**	**Phase 4**
Pilote 1	n = 700 $\overline{VH_{lb}}$ = - .4145	n = 2137 $\overline{EM}$ = .0125	n = 2642 $\overline{EM}$ = .0203	n = 5251 $\overline{EM}$ = .0112	n = 3883 $\overline{EM}$ = .0861
Pilote 2	n = 792 $\overline{VH_{lb}}$ = - .3697	n = 507 $\overline{EM}$ = .0212	n = 875 $\overline{EM}$ = .0952	n = 2257 $\overline{EM}$ = .0223	n = 7225 $\overline{EM}$ = .0428
Pilote 3	n = 2350 $\overline{VH_{lb}}$ = - .2535	n = 1164 $\overline{EM}$ = .0735	n = 1468 $\overline{EM}$ = .1609	n = 2797 $\overline{EM}$ = .0579	n = 9049 $\overline{EM}$ = - .0066

TABLE 2 – Valeur harmonique moyenne de la ligne de base de chaque pilote et écarts mélodiques moyens des phases 1 à 4 par pilote

4 Discussion

Dans cette étude, nous avons recueilli de la parole spontanée en situation simulée disposant d'une bonne validité écologique (simulateur reproduisant fidèlement un cockpit, scénario crédible, etc.). La parole avait un rôle majoritairement fonctionnel (communications radio). Son analyse s'est révélée riche : le recours au calcul de l'écart mélodique a permis d'observer l'évolution de la fréquence fondamentale en comparant chaque pilote à lui-même et aux autres lors des différentes phases de vol. Un effet d'interaction a été mis en avant : les trois sujets voient leur fréquence fondamentale augmenter sous le poids du stress engendré par la panne générale d'alimentation électrique. Par contre, celle-ci varie de manière contrastée entre les pilotes lors de l'atterrissage : la résistance au stress acquise lors de la formation initiale (Master) du pilote 3 explique peut-être cette disparité.

Bien que la corrélation entre l'écart mélodique et l'hétéro-évaluation soit faible, elle est plus importante que celle obtenue par Huttunen *et al.* (2011) et conforte l'idée que la fréquence fondamentale est un bon candidat pour estimer la charge cognitive et le stress qu'elle induit (Harmegnies & Landercy, 1992; Ruiz *et al.*, 1996; Giddens *et al.*, 2013). Toutefois, un travail important reste à mener pour caractériser le stress induit par la charge cognitive en fonction de ses trois composantes (la tâche, l'individu et l'environnement) et de leurs interactions. En effet, au cours de la simulation, nous avons pointé des événements susceptibles d'agir sur le pilote du fait de leur valence émotionnelle (évitement du missile, panne dans le noir complet, endossement du rôle de leader, etc.) et d'autres impactant plutôt les processus cognitifs requis pour l'effectuation de la tâche (check-list pour l'appréciation des avaries et pour le redémarrage du générateur principal, planification de l'atterrissage, approche aux instruments, etc.). Les répercussions de ces événements varient bien sûr en fonction de l'individu, c'est pourquoi il est crucial de collecter un maximum d'informations sur celui-ci.

Notons par ailleurs que l'évolution dans le temps de l'effet des phénomènes devrait être questionnée : nos données quant à l'auto-évaluation et à l'anxiété état suggèrent un effet cumulatif. Les valeurs croissent progressivement sans jamais diminuer. Ce cumul joue très probablement un rôle dans l'apparition de la fatigue et, ipso facto, peut entraîner une réduction des ressources cognitives.

Dans cette expérience, nous avons délibérément adopté la même démarche que Huttunen *et al.* (2011) en découpant le scénario en plusieurs phases. Toutefois, en prenant un peu de recul, nous nous interrogeons sur la pertinence de ce choix : en effet, les phases créées ont des longueurs dissemblables et variables d'un pilote à l'autre (voir la variabilité du nombre de segments voisés analysés dans la

table 2) et ne génèrent pas un niveau constant de charge cognitive. Nous suspectons que la moyenne de l'écart mélodique par phase gomme trop les variations au sein de chaque phase. La fréquence fondamentale changeant très rapidement (Harmegnies & Landercy, 1992; Ruiz *et al.*, 1996; Giddens *et al.*, 2013), une approche par événements et par épisodes nous paraît sans aucun doute plus indiquée, c'est pourquoi nous avons placé les faits les plus marquants sur la ligne du temps modélisant le scénario (figure 1). Il convient de signaler que chaque incident n'est pas systématiquement suivi d'une manifestation orale, ce qui rend parfois impossible la spéculation sur les effets de la charge cognitive via l'analyse du signal de parole. De plus, la réaction émotionnelle à un événement peut être différée dans le temps et avoir des effets plus ou moins durables. Par ailleurs, dans le domaine militaire, l'entraînement intensif vise, entre autres, à améliorer la régulation des réactions émotionnelles suscitées par l'émergence d'événements imprévus.

En outre, nous avons observé un fait particulier : les pilotes, francophones, communiquent en anglais avec la tour et leurs alliés mais, lorsqu'ils se parlent à eux-mêmes, le font en français même lorsqu'ils emploient une terminologie anglophone. Ce constat est sans doute à mettre en lien avec la compétence des sujets en langue anglaise (Piccaluga *et al.*, 2007) et l'adoption de la langue maternelle est sans doute une stratégie amoindrissant la charge cognitive (Martin *et al.*, 2013), ce qui suggère, a contrario, que l'utilisation de l'anglais est consommatrice de ressources cognitives.

Ces constatations montrent toute l'importance du contexte dans lequel chaque événement se déroule (Chanquoy *et al.*, 2007) et nous amènent à envisager la construction d'une typologie des situations en fonction de la durée de chaque événement, de sa valence émotionnelle, de la langue employée, de l'intention de communication du locuteur et de la nature précise de la tâche. A propos de cette dernière, notre étude confirme la pertinence des trois dimensions (conscience de la situation, charges informationnelle et décisionnelle) mises en avant par Huttunen *et al.* (2011).

Outre la fréquence fondamentale, les analyses spectrales sont susceptibles d'apporter des éléments intéressants pour étudier les effets de la charge cognitive. Nous aborderons nos futures recherches sous l'angle de la collaboration avec des chercheurs issus d'autres disciplines scientifiques. C'est là tout l'intérêt du projet BIOVOC dans lequel nous nous sommes engagés : la validation croisée des indices acoustiques avec des informations issues de paramètres physiologiques tels que la fréquence cardiaque (Alvear *et al.*, 2013) et biologiques recueillis au travers de la salive par exemple, nous offrira un regard nouveau qui nous permettra, nous l'espérons, de mieux définir et opérationnaliser les concepts impliqués.

Remerciements

Nous tenons à remercier les pilotes, le pilote instructeur et la force aérienne belge.

Références

ALVEAR R. M. B. D., BARÓN-LÓPEZ F. J., ALGUACIL M. D. & DAWID-MILNER M. S. (2013). Interactions between voice fundamental frequency and cardiovascular parameters. Preliminary results and physiological mechanisms. *Logopedics Phoniatrics Vocology*, **38**(2), 52–58.

BARROUILLET P., BERNARDIN S. & CAMOS V. (2004). Time Constraints and Resource Sharing in Adults' Working Memory Spans. *Journal of Experimental Psychology : General*, **133**(1), 83–100.

CHANQUOY L., TRICOT A. & SWELLER J. (2007). *La charge cognitive : théorie et applications*. Paris : Armand Colin.

ENDLER N. S. & PARKER J. D. A. (1990). *Coping Inventory for Stressful Situations (CISS) : Manual*. Toronto , Canada : Multi-Health Systems.

ENDSLEY M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors : The Journal of the Human Factors and Ergonomics Society*, **37**(1), 32–64.

GIDDENS C. L., BARRON K. W., BYRD-CRAVEN J., CLARK K. F. & WINTER A. S. (2013). Vocal indices of stress : A review. *Journal of Voice*, **27**(3), 390.e21 – 390.e29.

HANSEN J., SWAIL C., SOUTH A., MOORE H., STEENEKEN H., CUPPLES E., ANDERSON T., VLOEBERGHS C., TRANCOSO I. & VERLINDE P. (2000). *The impact of speech under "stress" on military speech technology = l'impact de la parole en condition de "stress" sur les technologies vocales militaires*. Neuilly-sur-Seine Cedex, France : RTO/NATO.

HARMEGNIES B. & LANDERCY A. (1992). A multivariate approach for the analysis of speech under cognitive stress. In *Proceedings of the ESCA Workshop "Speech Processing in Adverse Conditions*, p. 231–234, Nice.

HUTTUNEN K., KERÄNEN H., VÄYRYNEN E., PÄÄKKÖNEN R. & LEINO T. (2011). Effect of cognitive load on speech prosody in aviation : Evidence from military simulator flights. *Applied Ergonomics*, **42**(2), 348–357.

JOHANNES B., SALNITSKI V. P., GUNGA H. C. & KIRSCH K. (2000). Voice stress monitoring in space–possibilities and limits. *Aviation, Space, and Environmental Medicine*, **71**(9 Suppl), A58–65.

MARTIN C., HOURLIER S. & CEGARRA J. (2013). La charge mentale de travail : un concept qui reste indispensable, l'exemple de l'aéronautique. *Le travail humain*, **76**(4), 285–308.

MENDOZA E. & CARBALLO G. (1998). Acoustic analysis of induced vocal stressby means of cognitive workload tasks. *Journal of Voice*, **12**(3), 263–273.

PICCALUGA M., NESPOULOUS J.-L. & HARMEGNIES B. (2007). Disfluency surface markers and cognitive processing ; the case of simultaneous interpreting. In *16h International Congress of the Phonetic Sciences*, p. 1317–1320, Saarbrucken, Allemagne.

ROTHKRANTZ L. J. M., WIGGERS P., VAN WEES J.-W. A. & VAN VARK R. J. (2004). Voice Stress Analysis. In D. HUTCHISON, T. KANADE, J. KITTLER, J. M. KLEINBERG, F. MATTERN, J. C. MITCHELL, M. NAOR, O. NIERSTRASZ, C. PANDU RANGAN, B. STEFFEN, M. SUDAN, D. TERZOPOULOS, D. TYGAR, M. Y. VARDI, G. WEIKUM, P. SOJKA, I. KOPEČEK & K. PALA, Eds., *Text, Speech and Dialogue*, volume 3206, p. 449–456. Berlin, Heidelberg : Springer Berlin Heidelberg.

RUIZ R., ABSIL E., HARMEGNIES B., LEGROS C. & POCH D. (1996). Time- and spectrum-related variabilities in stressed speech under laboratory and real conditions. *Speech Communication*, **20**(1–2), 111 – 129. Speech under Stress.

SPIELBERGER C. D., BRUCHON-SCHWEITZER M. & PAULHAN I. (1993). *Inventaire d'anxiété état-trait forme Y (STAI-Y) : [manuel]*. Paris : Ed. du Centre de psychologie appliquée.

VAN LIERDE K., VAN HEULE S., DE LEY S., MERTENS E. & CLAEYS S. (2009). Effect of psychological stress on female vocal quality. A multiparameter approach. *Folia phoniatrica et logopaedica : official organ of the International Association of Logopedics and Phoniatrics (IALP)*, **61**(2), 105–111.

ZWICKER E. & FELDTKELLER R. (1981). *Psychoacoustique : l'oreille, récepteur d'information*. Collection technique et scientifique des télécommunications. Masson.

Structures prosodiques des langues romanes

Philippe Martin
LLF, UFRL, Université Paris Diderot Sorbonne Paris Cité
Place Paul Ricœur, 75013 Paris, France
philippe.martin@linguist.univ-paris-diderot.fr

RESUME

La description phonologique de la structure prosodique des langues romanes apparait similaire lorsque les interactions entre les accents mélodiques est prise en compte (ce qui n'est pas le cas dans la théorie autosegmentale-métrique). L'analyse acoustique de plus de 2600 énoncés lus et spontanés suggère que la réalisation des accents mélodiques, décrits en termes de contours mélodiques plutôt que de cibles tonales, indiquent avec les contours de frontière, des relations de dépendance « vers la droite » entre groupes accentuels. Ces relations permettent par incrémentation successive dans l'axe du temps la reconstitution par l'auditeur de la structure prosodique voulue par le locuteur. Dans ce cadre théorique, les langues romanes (italien, espagnol, catalan, portugais, roumain) utilisent les mêmes contours phonologiques pour indiquer les relations de dépendance menant au codage de la structure prosodique. Le français, dépourvu d'accent lexical, utilise un système de contours différent.

ABSTRACT

Prosodic Structures of Romance Languages

Prosodic phonologic description of Romance Languages (except French) appears surprisingly similar, once the interaction between pitch accents as melodic contours rather than tonal targets is considered. Acoustic analysis performed on more than 2,600 sentences of both read and spontaneous speech suggests that the realizations of pitch accent, together with boundary tones, do indicate a relation of dependency from one accent phrase towards another accent phrase "on its right" (i.e. occurring in the future of the sentence). This process eventually leads to a reconstitution by the listener of the prosodic structure intended by the speaker through an incremental process along the time scale. French, deprived from lexical stress, uses another set of melodic contours.

MOTS-CLES : intonation, langues romanes, français, structure prosodique.

KEYWORDS: sentence intonation, Romance language, French, incremental prosodic structure.

1. Introduction

La comparaison des caractéristiques prosodiques des langues romanes a fait l'objet de plusieurs études depuis quelque dix ou quinze ans. Le projet AMPER par exemple (Contini et al., 2002) décrit les différences prosodiques de phrases comparables dans de nombreuses variétés de langues romanes.

Plus récemment, dans un ouvrage édité par S. Fróta et P. Prieto (2015), des différences prosodiques détaillées ainsi que les variations liées à la modalité sont analysées et comparées non seulement pour les langues romanes nationales (espagnol et portugais européen, italien, roumain, français), mais aussi pour des réalisations régionales comme le catalan, le friulien, l'occitan et le sarde. Le cadre théorique adopté est autosegmental-métrique, et la transcription des données utilise le système ToBI.

On relate ici un travail similaire, mais qui utilise un cadre théorique totalement différent dans le but de mieux différentier les traits phonologiques des détails phonétiques dans les réalisations d'énoncés lus et spontanés pour six langues romanes, le français, l'italien, le catalan, le roumain ainsi que l'espagnol et le portugais européens. De cette approche résulte une grande similitude phonologique dans l'indication de la structure prosodique des langues considérées, à l'exception du français. En effet, le français, dépourvu d'accent lexical, utilise un système différent pour indiquer la structure prosodique des énoncés.

2. Cadre théorique

Alors que l'approche autosegmentale-métrique rend compte de l'intonation de l'énoncé en collectant des séquences tonales bien formées telles que transcrites avec le système de notation ToBI, le cadre théorique utilisé ici est focalisé sur les caractéristiques des évènements prosodiques considérés dans leur séquence temporelle. Au lieu d'être envisagée comme un objet statique, la structure prosodique est décrite du point de vue phonologique comme un processus dynamique dans lequel les contours mélodiques (i.e. les variations de hauteur) à l'endroit des syllabes accentuées et en fin de groupe prosodique assurent par des relations de dépendance entre contours la construction incrémentale de la structure prosodique voulue par le locuteur et reconstruite par l'auditeur.

Une relation de dépendance entre deux objets phonologiques s'établit lorsque l'occurrence d'un des objets ne peut survenir que si l'autre objet est ou sera présent dans l'énoncé. Ainsi par exemple, dans le cas d'un processus dynamique comme l'élaboration par le locuteur et la reconstitution par l'auditeur de la structure prosodique, un contour dit de continuation majeure présuppose la réalisation future d'un contour terminal terminant l'énoncé. C'est donc le locuteur, dans la planification de la structure prosodique, qui tient compte dans la réalisation des contours successifs marquant la structure prosodique de l'existence dans le futur immédiat de contours dont ils dépendent, c'est à dire qu'un contour ne peut apparaitre que si la réalisation de celui dont il dépend est planifiée par le locuteur dans le déroulement de l'énoncé.

Il en résulte que les contours mélodiques indiquent nécessairement une relation de dépendance envers un autre contour situé « à droite », c'est-à-dire dans le futur du signal de parole, à l'exception de contour terminal conclusif, qui constitue la racine de la structure prosodique.

On peut montrer (Martin, 2009, 2015) que ces relations de dépendance définissent de manière incrémentale des regroupements successifs d'unités prosodiques minimales constituées par les groupes accentuels. Ces regroupements constituent la structure prosodique de l'énoncé. Un groupe accentuel, équivalent du *Accent Phrase* de la théorie Autosegmentale-Métrique, est constitué d'une séquence de syllabes dont une seule est accentuée, hors effet d'insistance.

Pour les langues romanes à l'exception du français, la syllabe accentuée correspond à l'accent lexical. Pour le français, elle correspond à la syllabe finale des groupes accentuels, dont la durée d'énonciation est limitée à quelque 1250 ms (Martin, 2015). Un exemple de regroupements successifs de groupes accentuels tels qu'indiqués par les contours mélodiques qui les terminent est donné Fig. 1. Il n'est peut-être pas inutile de rappeler que les contours prosodiques apparaissent séquentiellement dans le temps, et ne sont pas perçus simultanément par l'auditeur comme pourrait le suggérer la représentation planaire de la Fig. 1.

Pour que ce processus fonctionne, il faut (et il suffit) non seulement que les contours prosodiques soient différentiés, c'est-à-dire qu'il existe dans leurs réalisations une différentiation acoustique utilisant des traits de montée, de descente et/ou d'empan mélodique, mais aussi qu'ils soient ordonnés, c'est-à-dire qu'on puisse établir des relations de dépendance entre eux. L'observation des données mène à définir différentes classes de contours prosodiques nécessaires et suffisantes pour assurer l'indication des relations de dépendance entre groupes accentuels, classes

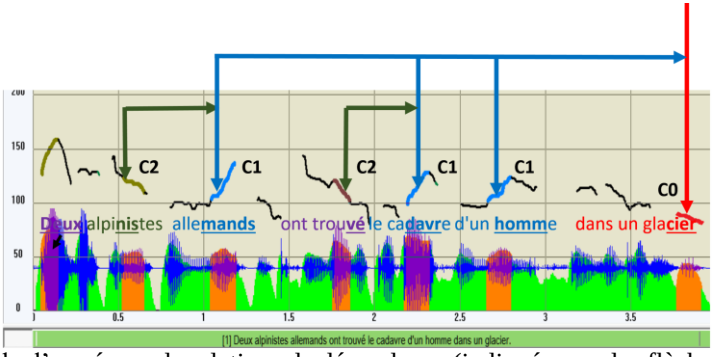


FIGURE 1 : Exemple d'un réseau de relations de dépendance (indiquées par les flèches horizontales) liées aux contours prosodiques définissant la structure prosodique de l'énoncé. On a donc C2 dépendant de C1, et C1 dépendant de C0. Le parenthésage dérivé [*deux alpi**nist**es C2 alle**mand**s C1*][*ont trou**vé** C2 le ca**dav**re C1*][*d'un **homm**e C1*][*dans un gla**cier** C0*] indique les regroupements successifs des groupes accentuels. On peut noter dans cet exemple lu que la structure prosodique n'est pas totalement congruente avec la structure syntaxique (*deux alpi**nist**es alle**mand**s*) (*ont trou**vé*** (*le ca**dav**re d'un **homm**e*)) (*dans un gla**cier***).

instanciées par des variations (contours) mélodiques placés sur les syllabes accentuées (et essentiellement sur leur noyau vocalique) :

C0 : contour terminal conclusif (cas déclaratif), descendant et bas.

Cc : contour mélodique complexe, plat ou descendant sur la syllabe accentuée du groupe accentuel, et montant sur la syllabe finale. Si cette dernière est accentuée, les deux mouvements successifs descendant et montant sont réalisés sur cette même syllabe (n'existe pas en français).

C1 : contour montant perçu comme tel et non comme un ton statique, c'est-à-dire phonétiquement supérieur au seuil de glissando (Rossi, 1971).

C2 : contour descendant perçu comme tel et non comme un ton statique, c'est-à-dire phonétiquement supérieur au seuil de glissando.

Cn : contour neutralisé, descendant, plat ou montant perçu comme un ton statique (donc dont la variation mélodique est phonétiquement inférieure au seuil de glissando).

Le seuil de glissando est défini à partir des valeurs de fréquence fondamentale de début et de fin de voyelle des syllabes accentuées. Bien que ce seuil ait été établi à l'origine sur la perception de voyelles synthétiques, et qu'il implique, mais dans une moindre mesure, d'autres paramètres que la vitesse de variation mélodique, il constitue un indice parmi d'autres pour différentier à partir de leurs réalisations phonétiques les contours C1 et C2 du contour neutralisé Cn.

En établissant par l'analyse des données expérimentales les relations de dépendance transitives entre ces contours prosodiques, on obtient, pour le français, dépourvu d'accent lexical et donc de contour complexe Cc :

Cn -> C2 -> C1 -> C0 (dépendance de Cn par rapport à C2, de C2 par rapport à C1, de C1 par rapport à Co le contour terminal de fin d'énoncé).

et pour les autres langues romanes :

Cn -> C1 -> C2 -> Cc -> C0 (dépendance de Cn par rapport à C1, de C2 par rapport à C2, de C2 par rapport à Cc, de Cc par rapport à C0 le contour terminal de fin d'énoncé).

La hiérarchie des classes de contours est alors, pour les langues romanes autres que le français Cn < C1 < C2 < Cc < C0, et pour le français, dépourvu d'accent lexical et donc de contour complexe Cc,

Cn < C2 < C1 < C0. La hiérarchie des contours montant C1 et descendant C2 est donc inversée pour le français par rapport aux autres langues romanes.

En comparant deux contours mélodiques successifs (donc placés sur deux syllabes accentuées successives ne constituant pas d'accents d'insistance), et tenant compte de leur ordonnancement, les regroupements de groupes accentuels s'opèrent à parti des trois règles suivantes :

Soient Cx et Cy des contours appartenant à l'ensemble C0, Cc, C1, C2, Cn,

Si Cx < Cy structure partielle [Cx Cy] les groupes prosodiques dont le dernier groupe accentuel contient les contours Cx et Cy forment un groupe plus grand. Par exemple la séquence C2 Cc indiquent une dépendance de C2 relativement à Cc et déterminent le regroupement des unités prosodiques (groupes accentuels ou regroupement de groupes accentuels) porteurs de C2 et Cc.

Si Cx = Cy structure partielle [Cx Cy…] les groupes prosodiques dont le dernier groupe accentuel contient les contours Cx et Cy font partie d'une liste, terminée plus tard par l'apparition d'un contour de rang plus élevé. Ainsi, la séquence C1 C1 forme une liste partielle terminée par un contour C2 ou Cc (pour les langues romanes) et par C0 pour le français.

Si Cx > Cy structure partielle [Cx [Cy…les groupes prosodiques dont le dernier groupe accentuel contient les contours Cx et Cy ne forment pas un groupe plus grand. Une séquence de contours telle que Cc C2 ne détermine donc pas un groupe prosodique plus grand dont les syllabes accentuées de l'unité accentuelle finale seraient Cc et C2.

Ce processus est donc local, et implique également la transitivité. Ainsi le contour neutralisé Cn peut indiquer une relation de dépendance aussi bien avec C1, C2, Cc ou C0, puisque ces contours sont situés à un rang supérieur dans la hiérarchie des contours. Ce n'est que quand la complexité relative de la structure prosodique le requiert que des contours de rang immédiatement inférieur sont obligatoirement réalisés par le locuteur, sinon ils sont optionnels.

Comme signalé plus haut, les relations de dépendance opèrent « à droite » relativement au futur de la séquence de contours mélodiques. Ainsi un contour C1 de continuation majeure, en indiquant non seulement que l'énoncé n'est pas terminé, présuppose l'apparition d'un contour terminal C0 signalant la fin de l'énoncé et permettant à l'auditeur de rassembler les différents groupements de syntagmes pour accéder au sens de l'énoncé. L'existence de C1 dépend donc de l'apparition future d'un contour C0. Il en va de même pour le contour descendant C2, qui dépend de l'apparition future d'un contour C1, etc.

3. Analyse expérimentale

Le corpus d'analyse comporte :

1. Un ensemble de 60 phrases de complexité syntaxique croissante lues par deux locuteurs, extrait du corpus EuRom4 (1991-1997). Les phrases à l'origine en français, italien, espagnol et portugais européen et ont été adaptées en catalan et en roumain.

2. Le corpus EuRom4 (français, italien, espagnol, portugais européen), contenant 24 histoires courtes de longueur variable lues par 4 locuteurs dans chaque langue. Le nombre moyen de phrases pour chaque langue est d'environ 200.

3. Le corpus EuRom5 (Bonvino 2011) français, italien, espagnol, catalan, portugais européen, composé de 20 histoires courtes lues par deux locuteurs professionnels, avec environ 300 phrases pour chaque langue.

L'analyse acoustique a été réalisée avec le logiciel WinPitch (1996-2016). Les transcriptions de texte étant préalablement disponibles pour les trois corpus, l'alignement texte-parole a été réalisée la fonction d'alignement de volée, permettant un alignement rapide et efficace des transcriptions de texte. Cet alignement permet l'analyse des informations prosodiques retrouvées en un clic à partir de

n'importe quelle partie du texte. Une toute petite partie des résultats est présentée ci-dessous, afin d'illustrer quelques aspects de la procédure de découverte des relations de dépendance entre contours. Des données plus complètes sont disponibles dans Martin (2015).

Les figures qui suivent présentent des courbes mélodiques dont les segments continus plus épais correspondent aux voyelles accentuées et donc aux contours C0, C1 et C2. Le contour Cn est représenté en trait plus fin, et les traits interrompus correspondent à la variation mélodique montante de frontière relative aux contours Cc.

Comment alors déterminer la structure prosodique d'un exemple donné, celle-ci étant par hypothèse indépendante des autres structures de l'énoncé, et en particulier de la structure syntaxique ? Ayant posé que les groupes accentuels en constituent les unités minimales, on procède par expansions successives du nombre de ces groupes accentuels dans l'analyse des données. Ainsi avec un seul groupe accentuel on ne peut former qu'un seule structure prosodique, terminée par un contour terminal conclusif C0 indiquant également la modalité déclarative ou interrogative (et ses variantes) de l'énoncé (on se limite ici à la modalité déclarative).

Deux groupes accentuels peuvent constituer deux structures prosodiques distinctes, indiquées par la séquence de contours C0 Cx et Cy C0, Cx et Cy devant être déterminés. On a donc soit un contour terminal C0 identifié sur le premier groupe accentuel, soit sur le second, soit encore sur les deux (cas d'une epexégèse). L'identification d'un contour terminal se fait par test de perception de la finalité en segmentant éventuellement l'énoncé avec un éditeur de signal. On trouve alors que dans le premier cas C0 Cx, le contour Cx est plat et sera noté Cn (configuration classique dite *propos-thème* ou *noyau-postnoyau*).

Dans la seconde configuration Cy C0, le contour Cy peut être réalisé de plusieurs manières que l'on peur recenser dans les données : Cc (contour complexe), C1 (continuation majeure) ou Cn (contour neutralisé). Puisqu'un seul trait acoustique suffit à différentier le contour Cy de C0, le locuteur peut choisir parmi les contours disponibles Cn, ou utiliser plus de traits acoustiques de différentiation avec C1 ou Cc pour des raisons stylistiques par exemple.

Trois structures prosodiques distinctes peuvent hiérarchiser trois groupes accentuels A, B et C : [A] [B] [C], [A B] [C] et [A] [B C]. Pour s'assurer de la réalisation effective d'une de ces structures dans un exemple donné, et ainsi déterminer les contours qui l'indiquent, on peut utiliser des énoncés dépourvus de structure syntaxique et informationnelle, telle que les énumérations, les numéros de téléphone, etc.

C'est le cas de la Fig. 2, montrant la dépendance en italien de C2 par rapport à Cc, et de C1 par rapport à C0. Tous les groupes sont structurés avec le contours C2 Cc, sauf le dernier par C1 C0, illustrant les relations de dépendance C2 - > Cc et C1 - > C0. (de Italo Svevo, La Coscienza Di Zeno, lu par Moro Silo, Il narratore, 2006).

Italien [***B*** *C2 antiparasi**ta**ri Cc*] [***C*** *C2 anticoncezio**na**li Cc*]…[*M C1 an**ti**lopi C0*].

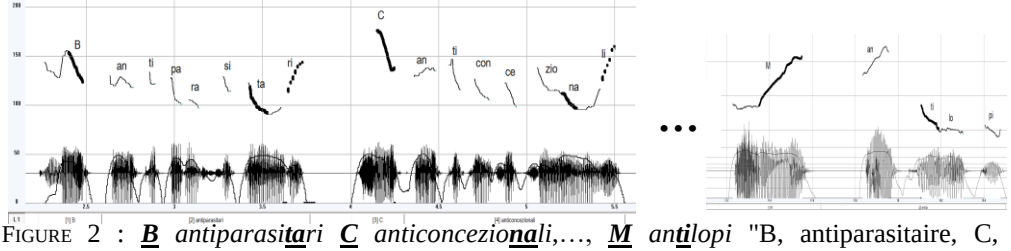


FIGURE 2 : ***B*** *antiparasi**ta**ri* ***C*** *anticoncezio**na**li,…*, ***M*** *an**ti**lopi* "B, antiparasitaire, C, anticonceptionnel,…, M, antilope.", exemple extrait d'une longue énumération faite à partir d'une séquence de groupes de deux mots prosodiques, terminé par Cc, sauf le dernier se terminant par un contour C0 conclusif placé sur la syllabe accentuée *ti* de *antilopi*.

La configuration ci-dessus se retrouve non seulement en italien, mais aussi dans les autres langues romanes comme le montrent les figures 3 à 5.

Espagnol (Eurom5 E12-1) C2 Cc : [***cuan****do C2*] [*se cons****ta****te Cc*]

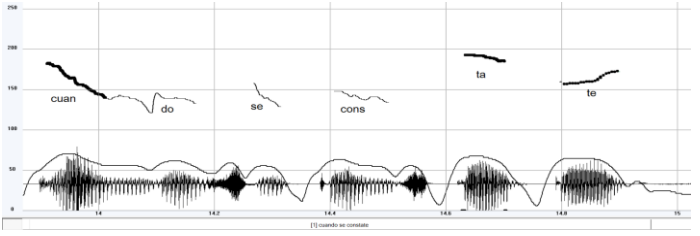


FIGURE 3. *une pide asus 47 Estados miembros que establezcan algún tipo de sanción* ***cuan****do se cons****ta****te que hay una "provocación a la discriminación" en los mensajes publicitarios* "Une demande à ses 47 Etats membres de mettre en place tout type de sanction lorsqu'il est constaté qu'il existe une "incitation à la discrimination" dans la publicité".

Portuguais (Eurom5 P04-1) C2 Cc : [*apeli****da****do C2*] [*de* ***Ós****car Cc*] "appelé Oscar"

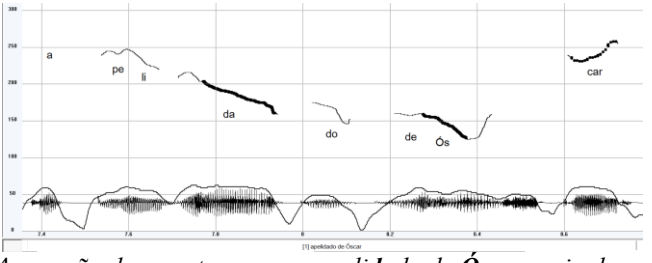


FIGURE 4. *um cão de raça terra nova apeli****da****do de* ***Ós****car, cujo dono é um socialite, vai ser submetido a um lifting aos olhos* "un chien de race terre-neuve surnommé Oscar, dont le propriétaire est un mondain, va subir un lifting des paupières"

Catalan (Eurom5 C19-1) C2 Cc : [***Com*** *C2*] [*les for****mi****gues Cc*]

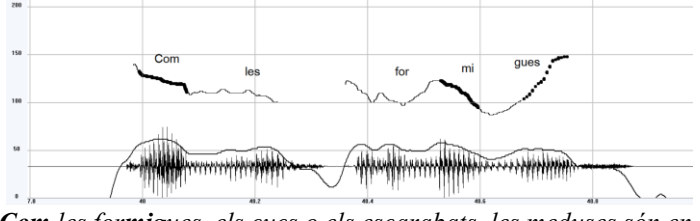


FIGURE 5. ***Com*** *les for****mi****gues, els cucs o els escarabats, les meduses són en algunes cultures un element bàsic de l'alimentació.* "Comme les fourmis, les coléoptères et les vers, les méduses sont dans certaines cultures un aliment de base".

Tous ces exemples illustrent la relation de dépendance C2 -> Cc et Cc –> C0, établie en partant de l'hypothèse d'une congruence au moins partielle portant sur les segments analysés des structures prosodiques et syntaxiques de surface. Cette congruence est généralement plus fréquente dans les phrases lues, le locuteur s'efforçant de faire correspondre les frontières prosodiques avec les frontières syntaxiques importantes dans la lecture. On note aussi que si ce sont les mêmes contours phonologiques, et donc les mêmes contrastes, qui sont utilisés dans ces langues romanes, leur réalisations phonétiques peut être différente ainsi que le montrent les tracés d'analyse acoustiques.

Portugais (Eurom5 P05-1)

[*é permi****ti****do C2*] [*ma****tar*** *C2*] [*um esco****cês*** *Cc*] "Il est permis du tuer un Écossais".

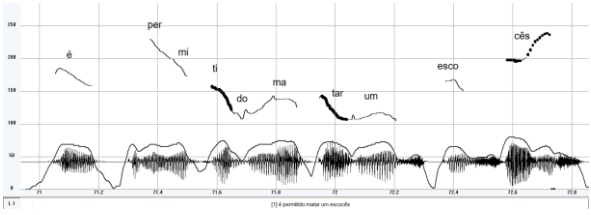


FIGURE 8. Courbe mélodique et contours de l'exemple *é permi**ti**do ma**tar** um esco**cês***.

Français

Le français ne possède pas de contour complexe Cc, et les relations de dépendance sont illustrés par les exemples des figures 6 et 7. L'exemple suivant illustre le contraste prototypique d'inversion de pente mélodique en français. [*les gar**çons** C2 de **pist**e C1*] (Eurom5 F03-1)

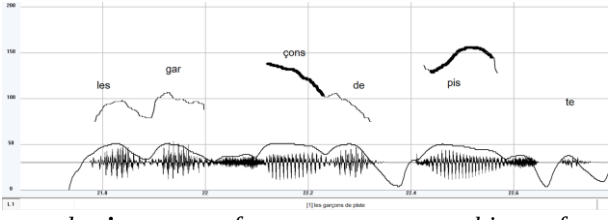


FIGURE 6. *Les gar**çons** de **pist**e se transforment en porteur, voltigeur, funambule ou cascadeur…*

Cet exemple illustre la relation de dépendance C2 -> C1, prototypique du français avec C1 -> C0. Les mêmes relations se trouvent dans des exemples plus complexes, comme illustré Fig. 7.

[*ain**si** C1*] [*sa nouvelle **gamm**e C2 de combi**né**s C2 présen**tés** Cn lun**di** C1*] **(Eurom5 F04-1)**

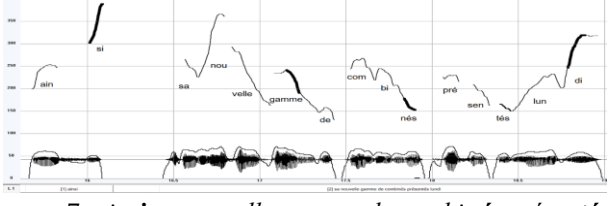
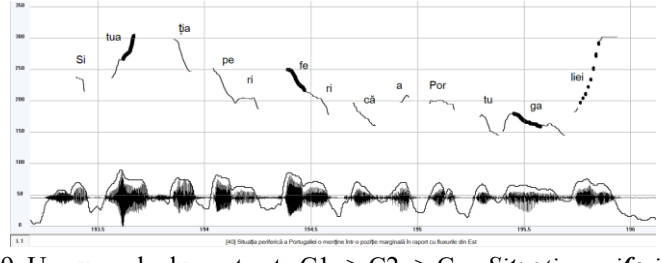


FIGURE 7. *ain**si** sa nouvelle **gamm**e de combi**né**s présen**tés** lun**di**.*

Les deux derniers exemples (Fig. 9 et 10) illustrent le fonctionnement des contours de relations de dépendance dans l'indication de deux regroupements différents II et III dans des lenagues romanex autrs que le français.

Roumain (Eurom4 rofn40)



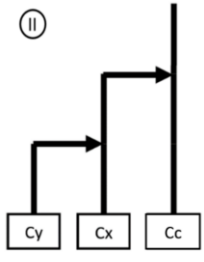


FIGURE 9. Un exemple du contraste C1 -> C2 -> Cc : *Si**tua**ția peri**fe**rică a Portu**ga**liei o menține într-o poziție marginală în raport cu fluxurile din Est*, réalisant la structure II.

Italien (Eurom4 I_23_06) [[[***che*** *C2*] [[*trasfe**rir**si C1*] [*in* ***U****SA Cc*]]] “lequel, transféré aux USA…”

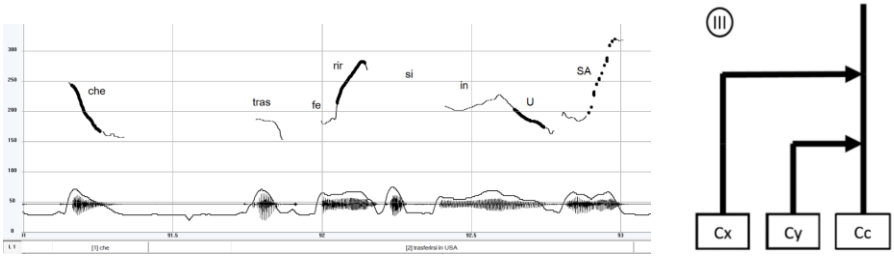


FIGURE 10. Un exemple du contraste C2 -> Cc et C1 -> Cc : ***che*** *trasfe**rir**si in* ***U****SA*

Ce dernier exemple est prototypique d'un contour mélodique C2 sur le mot monosyllabique *che* juste avant le début (la limite gauche) de la parenthèse *trasfe**rir**si in* ***U****SA*. Elle montre clairement que C2 fait contraste avec Cc et non avec le contour suivant C1, formant un groupe avec toute la parenthèse.

4. Conclusions

Le modèle de structure prosodique incrémentale est basé sur un principe clairement défini, l'indication d'une relation de dépendance "à droite" entre les événements prosodiques instanciés par des contours mélodiques des syllabes accentuées des groupes accentuels et en position finale de syntagmes intonatifs (cas du contour complexe Cc). À partir d'un inventaire des classes de contours mélodiques en termes de relations de dépendance, une description unifiée de la structure prosodique émerge pour toutes les langues romanes à accent lexical. Grâce à cette approche, la seule différence avec le français dépourvu d’accent lexical se rapporte à l'absence de contour complexe Cc, ce qui modifie le classement des contours mélodiques en préservant le principe de l'inversion de la pente mélodique.

En effet, dans les langues romanes le contraste de pente mélodique, à savoir montant -> descendant et descendant -> montant est appliqué dans la relation de dépendance C1 -> C2, C2 -> CC et CC -> C0, le contour complexe apparaissant phonologiquement comme une montée mélodique. En français, les relations existantes sont C2 -> C1 et C1 -> C0. Les langues romanes ont donc la possibilité de différencier un niveau supplémentaire dans la structure prosodique.

Cette conclusion va clairement à l’encontre du cadre théorique dominant basé sur la phonologie métrique-autosegmentale. En niant ex abrupto le rôle des accents mélodiques (pitch accents), cette approche ne peut évidemment rendre compte d’une quelconque interaction entre les contours mélodiques. De même, la définition des groupes accentuels n’a pas de rapport direct avec les proéminences métriques.

Une dernière remarque. Le cadre autosegmental-métrique ne propose en fait aucune explication quant à la fonction de la structure prosodique autre qu’actualiser la syntaxe, en se basant sur le concept du « bien-formé » emprunté à l’approche générative transformationnelle. Or on peut montrer (Martin, 2015) que la structure prosodique n’est pas dérivée de la syntaxe (si ce n’est dans une certaine mesure dans l’oralisation de l’écrit), elle la précède aussi bien pour le locuteur que pour l’auditeur. Du reste, il ne saurait y avoir de production langagière sans structure prosodique, même en lecture silencieuse, alors que l’inverse est possible.

Références

BONVINO E. et al. (2011). *EuRom5*, Ulrico Hoepli, Milano.

CONTINI M. et al. (2002). Un projet d'atlas multimédia prosodique de l'espace roman. *Proceedings of the 1st International Conference of Speech Prosody*, Aubenas d'Ardèche, Lienhart, 227-231.

EUROM4 (1991-1997). Projet européen Lingua (CEE) Institut National de la Langue Française (INALF), Responsable : Claire Blanche-Benveniste.

FROTA S., Prieto P. eds. (2015). *Intonation in Romance*. OUP Oxford, 400p.

MARTIN Ph. (2009) *Intonation du français*. Armand Colin, Paris, 256p.

MARTIN Ph. (2015) *The Structure of Spoken Language*. *Intonation in Romance*. Cambridge University Press, 340p.

ROSSI M. (1971). Le seuil de glissando ou seuil de perception des variations tonales pour la parole, *Phonetica* (23) 1-33.

WINPITCH. (2016). www.winpitch.com

Suivi de contours d'articulateurs orofaciaux à partir d'IRM dynamique

Mathieu Labrunie [1, 2] Pierre Badin [1, 2] Laurent Lamalle [3] Coriandre Vilain [1, 2]
Louis-Jean Boë [1, 2] Jens Frahm [4] Peter Birkholz [5]
(1) Univ. Grenoble Alpes, GIPSA-Lab, F-38000 Grenoble, France
(2) CNRS, GIPSA-Lab, F-38000 Grenoble, France
(3) Inserm US 17 — CNRS UMS 3552 — Univ. Grenoble Alpes & CHU de Grenoble UMS IRMaGE, France
(4) Biomedizinische NMR Forschungs GmbH am Max-Planck-Institut für biophysikalische Chemie, Göttingen, Germany
(5) Institute of Acoustics and Speech Communication, TU Dresden, Germany
(mathieu.labrunie, pierre.badin)@gipsa-lab.grenoble-inp.fr

RESUME

Nous présentons une méthode de prédiction de contours médiosagittaux des organes orofaciaux de la parole et la déglutition à partir d'images IRM dynamiques. Pour chaque locuteur, un ensemble de 60 images représentatives pour lesquelles les contours ont été tracés manuellement permet d'entraîner des modèles ACP d'images et de contours articulatoires, ainsi qu'un modèle multilinéaire qui prédit les paramètres des contours à partir des paramètres des images. Les contours obtenus sont ensuite corrigés par des modèles de forme actifs (ASM) modifiés utilisant les informations locales de profils d'intensité de pixels le long des normales aux contours. Les performances de cette méthode (erreurs moyennes « points à contour » entre 0,57 et 0,70 mm) sont insensibles au type de séquence IRM (écho de gradient avec échantillonnage synchronisé ou écho de gradient radial hautement sous-échantillonné), sont meilleures que celles de la littérature, et rendent possible le traitement de volumineux corpus d'images IRM dynamiques.

ABSTRACT

Orofacial articulators tracking from dynamic MRI.

We introduce a method for predicting midsagittal contours of orofacial organs during speech and swallowing from dynamic MRI images. For each speaker, a set of 60 representative images for which contours have been manually traced allows for training PCA images. The data serve to derive articulatory contour PCA models and a multilinear model that predicts contour parameters from image parameters. The obtained contours are then corrected by a modified Active Shape Model (ASM) using the local information of the pixel intensity profiles along the contour normals. The performance of this method (mean "points to contour" errors between 0.57 and 0.70 mm) is insensitive to the type of MRI sequence (conventional gradient echoes with synchronized sampling or highly undersampled radial gradient echoes), better than those in the literature, and make it possible to process large corpora of dynamic MRI images.

MOTS-CLES : IRM dynamique ; articulateurs orofaciaux de la parole ; suivi automatique de contours ; régression linéaire multiple ; modèles de forme actifs.

KEYWORDS: Dynamic MRI; speech orofacial articulators; automatic contour tracking; multiple linear regression; Active Shape Models.

1 Introduction

Les progrès considérables accomplis en IRM dynamique temps réel dans la dernière décennie (*cf.* Uecker *et al.* (2010) ou Niebergall *et al.* (2013)) ont rendu cette technique d'imagerie médicale extrêmement intéressante pour l'étude des mouvements des articulateurs orofaciaux dans les tâches de parole (Silva & Teixeira (2015)) ou de déglutition (Olthoff *et al.* (2014)), en permettant l'acquisition de volumineux corpus d'images médiosagittales à 30-60 images par seconde avec une résolution de l'ordre de 1,5 mm/pixel. Afin de pouvoir caractériser et modéliser ces données, il est donc nécessaire de développer des méthodes automatiques de suivi des contours des articulateurs à partir de ces images fournissant des résultats aussi précis et fiables que les méthodes (semi-) manuelles traditionnelles (voir p. ex. Serrurier & Badin (2008)).

Cet article décrit notre approche pour développer une telle méthode pour tous les organes articulatoires orofaciaux impliqués dans la production de la parole et la déglutition. Deux types de structures ont été considérés : les structures osseuses rigides, et les organes déformables. Les structures rigides du crâne doivent être suivies pour contrôler les mouvements de tête involontaires des sujets. Les autres structures rigides intéressantes sont la mâchoire et l'os hyoïde qui ont des mouvements spécifiques aussi bien en parole qu'en déglutition. Les organes déformables sont les lèvres supérieure et inférieure, la langue, l'épiglotte, le voile du palais, et l'ensemble de la paroi naso- et oro-pharyngée postérieure.

2 Travaux précédents

Avant de décrire notre approche, nous présentons une revue de la littérature sur le suivi d'articulateurs à partir d'IRM basée sur le travail très détaillé de Silva & Teixeira (2015).

Bresch & Narayanan (2009) proposent une méthode d'ajustement de contours dans l'espace de Fourier des images. Un modèle de contours est constitué des trois principales régions des organes articulatoires (au-dessus du palais dur, au-dessous de la langue et en arrière de la paroi pharyngée) délimitées par des frontières polygonales. Une descente de gradient minimise la distance entre l'image et les contours dans leurs espaces de Fourier. Bien qu'aucune évaluation quantitative ne soit donnée, les exemples de contours déterminés attestent d'une qualité raisonnable de la méthode.

Proctor *et al.* (2010) identifient les points de la ligne centrale du conduit vocal sur les images IRM en cherchant le chemin optimal entre les différentes positions de minima d'intensité de pixels le long de lignes d'un système de grilles perpendiculaires au conduit positionnées manuellement. Les frontières des tissus sont ensuite déterminées comme les positions de plus fort gradient d'intensité des pixels de chaque côté de la ligne centrale. Cette approche non supervisée fournit des contours de conduit indifférenciés, avec une erreur RMS de reconstruction de la distance sagittale variant de 0,82 à 1,61 mm. Kim *et al.* (2014) ont tenté d'améliorer cette méthode en optimisant la qualité d'image à l'aide d'une carte de correction de sensibilité de pixels et d'une réduction de bruit de grain par traitement d'image local, et un lissage des contours estimés. Ils obtiennent une erreur RMS sur la distance sagittale entre 2,13 et 2,79 mm avec des images de résolution de 3 mm/px.

Eryildirim & Berger (2011) ont développé un algorithme de segmentation de langue à partir d'images IRM statiques semblable aux algorithmes de type modèles de forme actifs (Active Shape

Models, ASM). Cet algorithme est basé sur un modèle de forme construit à partir des contours édités manuellement pour 38 images par analyse en composantes principales (ACP). La détection des points terminaux des contours de langue est améliorée par l'utilisation d'une méthode de recalage non rigide. Ils obtiennent une erreur moyenne de reconstruction des distances entre contours de 1,6 mm avec des images de résolution de 0,625 mm/px.

Finalement, Silva & Teixeira (2015) ont récemment proposé un modèle actif d'apparence modifié (Active Appearance Model, AAM) pour le suivi de contours articulatoires à partir d'images IRM dynamiques. Ils utilisent deux modèles AAM, l'un construit à partir des contours édités manuellement sur 30 images d'articulations non-nasales, et l'autre à partir de 21 articulations nasales. Ils trouvent que leur approche est plus rapide et converge mieux que les AAM traditionnels. Notons que chaque articulateur est clairement identifié à la fois lors de la segmentation manuelle et dans le modèle de forme (lèvres, corps et pointe de la langue, voile du palais, palais dur et pharynx). Ils mesurent les erreurs en termes du coefficient de similarité de Dice qui reflète la différence du nombre de pixels de part et d'autre des contours ; cette erreur n'est pas directement comparable à une distance RMS entre contours.

Nous introduisons dans cet article une méthode proche des ASM initialisée à l'aide d'une procédure de prédiction des contours à partir des intensités des pixels de l'image par modèle linéaire multiple. Nous décrivons l'implémentation de cette méthode et présentons des résultats d'évaluation sur des corpus d'images IRM de qualités différentes, ainsi que pour des méthodes basées sur le recalage d'images.

3 Segmentation basée sur le recalage d'image

Le recalage d'image consiste à déterminer la transformation spatiale – rigide ou élastique – qui permet de mettre en correspondance une image source avec une image cible. L'algorithme de recalage détermine la transformation optimale minimisant la distance, selon la mesure de similarité ou de dissimilarité choisie, entre les intensités des pixels de l'image source transformée et celles des pixels de l'image cible. Appliquer cette transformation de recalage à des contours de référence tracés sur l'image source permet ensuite de prédire les contours de l'image cible.

3.1 Recalage des structures rigides par comparaison à un motif standard

Pendant l'acquisition des données, la tête du sujet est stabilisée au mieux à l'aide de coussins en mousse, mais il est impossible d'empêcher complètement les mouvements parallèles au plan médiosagittal. Il est donc nécessaire de suivre les mouvements du crâne afin de les compenser. Par ailleurs, les mouvements d'autres structures rigides importantes pour l'articulation – la mâchoire et l'os hyoïde – devront être déterminés. La rigidité de ces structures n'autorise que des mouvements de translation et de rotation dans le plan médiosagittal. Une méthode de segmentation adaptée à ces propriétés rigides de l'objet d'intérêt est la comparaison à des motifs standards (*template matching*). La première étape de cette méthode consiste à choisir un motif contenant l'objet d'intérêt (p. ex. le palais) dont le contour est connu sur une image de référence. Ce motif est délimité par un masque excluant les tissus voisins non caractéristiques ou variables. Pour une image à traiter, l'objectif est ensuite de déterminer les paramètres de rotation et translation 2D pour lesquels le motif source de l'image de référence se superpose au motif cible correspondant sur l'image à traiter.

3.2 Recalage des organes déformables par démons

Pour les organes déformables tels la langue, les lèvres, ou le voile du palais il faut calculer un champ de transformation potentiellement non-linéaire pour pouvoir transformer l'image source en image cible. Nous avons testé le recalage par démons de Kroon & Slump (2009). Le champ de transformation associé à cette méthode est influencé par deux forces : une force interne dirigée par le gradient en chaque point de l'image, et une force externe dirigée par la différence entre intensités correspondantes de l'image source transformée et de l'image cible. Un facteur α permet de pondérer ces deux forces. Dans la procédure itérative de minimisation de la distance entre les intensités des images source et cible, il est possible de régulariser certains champs de déformation. A chaque itération, un champ de mise à jour du champ de transformation rajoute des déplacements au champ de transformation préalablement obtenu. Nous avons employé une régularisation fluide sur ce champ de mise à jour (filtrage gaussien d'écart-type σ_{fluide}) et un effet de diffusion en réalisant un filtrage gaussien du champ de transformation (écart-type : σ_{diff}). Les paramètres optimaux que nous avons trouvés pour l'ensemble de nos corpus sont $\alpha = 12$, $\sigma_{diff} = 1$, et $\sigma_{fluide} = 4$.

4 Segmentation basée sur des méthodes d'apprentissage

Dans la méthode précédente, seules les informations d'une image et d'un contour de référence sont prises en compte pour obtenir le contour associé à une nouvelle image. Les méthodes de recalage d'image, rigide ou élastique, ne prennent en compte que les propriétés de l'image, mais ignorent les propriétés des contours recherchés. Disposer de tracés manuels experts des contours d'intérêt sur un corpus représentatif de l'ensemble des données offre la possibilité d'introduire une information pertinente sur les contours recherchés qui permet d'améliorer considérablement les résultats en contraignant l'espace de recherche des contours. Nous décrivons ci-dessous trois méthodes utilisant l'entraînement de modèles à partir d'une base d'images et de contours associés, après avoir indiqué comment nous sélectionnons les images du corpus d'apprentissage.

4.1 Sélection du corpus d'apprentissage

Pour construire un modèle suffisamment général pour représenter toutes les articulations d'intérêt, le corpus d'apprentissage doit couvrir aussi exhaustivement que possible la diversité des articulations que peut produire le locuteur, tout en minimisant le nombre d'images dont les contours devront être édités manuellement. Pour construire cet ensemble, nous avons supposé que la distance euclidienne entre l'intensité des pixels des images était une métrique corrélée avec la distance euclidienne entre contours associés (a posteriori nous avons trouvé des corrélations supérieures à 0,85, ce qui valide cette hypothèse). Nous avons donc réparti toutes les images en n_{cl} classes par classification ascendante hiérarchique en utilisant cette métrique : différents tests ont montré que $n_{cl} = 60$ constitue un bon compromis entre le nombre d'images à tracer et un niveau d'erreur de l'ordre du millimètre, et qu'en outre cette métrique produit un dendrogramme cohérent au sens du coefficient de corrélation cophénétique. Le représentant de chaque classe est ensuite choisi comme l'élément de la classe le plus éloigné des éléments des autres classes, de façon à assurer que la périphérie de l'espace soit également bien représentée.

4.2 Régression linéaire multiple (Multiple Linear Regression, MLR)

Le modèle de prédiction des contours en fonction des images le plus simple est celui qui prédit chacune des coordonnées de chacun des contours comme combinaison linéaire des intensités des

pixels de la zone d'intérêt. Deux types de zones ont été utilisés : une zone cadrée globalement (*cf.* le cadre jaune en Fig. 1a), ou des zones cadrées sur chaque articulateur (*cf.* les autres cadres de Fig. 1a). Nous avons réduit la dimensionnalité de l'espace des intensités des zones par ACP, en retenant n_{int_gbl} composantes pour le cadrage global, et $n_{int\_orgs}(1{:}n_{org})$ pour les cadrages des n_{org} organes. De même, les coordonnées des contours des organes sont modélisées soit par un seul ensemble de n_{cnt_gbl} composantes, soit séparément par $n_{cnt\_orgs}(1{:}n_{org})$ composantes pour chaque organe. Les nombres de composantes sont choisis pour minimiser les erreurs de prédiction pour chaque méthode. La deuxième méthode apporte une flexibilité supplémentaire qui permet de mieux approcher chaque contour d'organe ; les composantes des organes peuvent alors être partiellement corrélées entre organes, comme par exemple pour la lèvre inférieure et la mâchoire. Le modèle d'association entre les contours et les images s'obtient finalement par régression linéaire multiple des prédicteurs des contours en fonction des prédicteurs des intensités sur l'ensemble des n_{cl} données d'apprentissage, soit de manière globale, soit organe par organe.

4.3 Modèles de forme actifs (Active Shape Models, ASM)

La méthode générale des ASM (Cootes *et al.* (1995)) vise à ajuster les points d'un contour aux limites d'un objet dans une image en les déplaçant de manière itérative afin de minimiser la distance entre l'apparence mesurée au voisinage de ces points (un profil d'intensité par exemple) et celle prédite par un *modèle d'apparence*, tout en contraignant le contour par un modèle (*modèle de forme*). Ces modèles sont établis lors d'une phase d'apprentissage à partir des images et des contours tracés. Dans notre implémentation[1], que nous appellerons ASM modifié (ASMM) nous utilisons les modèles de forme décrits en 4.2. Des modèles d'apparence sont développés pour chaque point de chaque organe, pour trois niveaux d'échantillonnage des images (échelles de 2, 1, et 0.5), de la manière suivante. En chaque point du contour considéré, un profil d'intensité est échantillonné par interpolation sur $n_{pfl} = 13$ points distribués le long d'un segment normal centré sur le point par pas d'un pixel (voir Fig. 1b, haut). Au lieu de modéliser l'apparence de ces profils d'intensité par ACP comme dans les ASM traditionnels, nous associons ces profils à des classes. Deux classes principales sont déterminées. La classe « non-contact » regroupe les profils pour lesquels la distance du point du contour aux organes voisins le long de la normale est supérieure à un seuil de 2 pixels. La classe « contact » regroupe tous les autres cas, y compris donc ceux pour lesquels la distance entre contours est inférieure au seuil de 2 pixels. Cependant une classification plus fine est nécessaire, car plusieurs sortes de profils peuvent être obtenues pour une même classe, du fait de la variabilité de l'orientation des normales et des niveaux de gris pour les tissus. Ces deux classes ont donc été divisées par un algorithme des k-moyennes en sous-classes dont le nombre a été optimisé (jusqu'à 10 en pratique). Chaque sous-classe est finalement représentée par son profil moyen (voir deux exemples à la Fig. 1c, haut).

La procédure de segmentation débute par une initialisation du contour de l'organe considéré. Pour chaque point de chaque contour, on explore l'apparence dans le voisinage (voir Fig. 1b, bas) en déterminant les profils d'intensité à n_{pfl} points le long de la normale en faisant varier la position i_{ctr} du centre par pas de 1 pixel sur une plage de ±4 pixels (voir Fig. 1c, bas). On calcule ensuite les distances de tous ces profils aux profils moyens de toutes les sous-classes. Si la distance minimale correspond à la classe « contact » il est impossible de déterminer avec précision la position du point de contour qui est alors ignoré dans l'étape suivante. Sinon, le point d'indice i_{ctr} est considéré comme le point de contour corrigé. L'étape suivante consiste à ajuster le modèle de forme aux

[1] Notre implémentation est basée sur le script fourni par Dirk J Kroon dans http://www.mathworks.com/matlabcentral/fileexchange/26706-active-shape-model—asm--and- activeappearance- model--aam-

points déterminés à l'étape précédente. Notons que les points non déterminés dans les zones de contact sont reconstruits par le modèle de forme lors de ce processus. Cette procédure est exécutée pour chacune des trois résolutions, de la plus grossière à la plus fine, le résultat de l'une servant d'initialisation à la suivante.

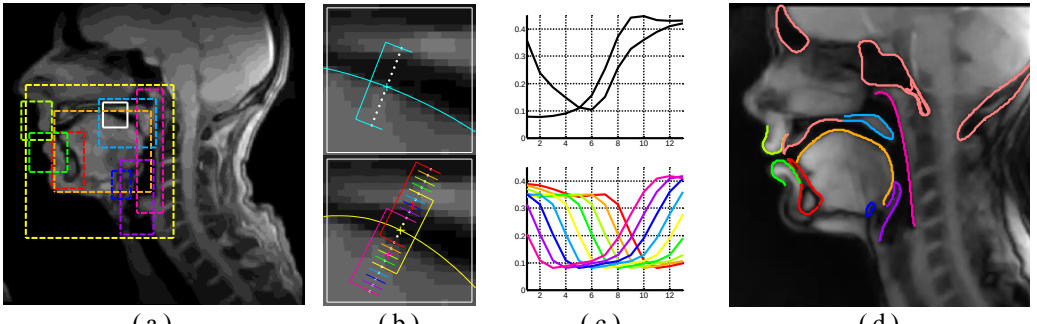
5 Evaluation

5.1 Données IRM pour l'évaluation

Nous avons testé les différentes méthodes présentées ci-dessus sur trois ensembles de séquences d'image IRM dynamiques obtenues par différentes techniques. Chaque ensemble était représenté par $n_{cl} = 60$ images sélectionnées par classification hiérarchique (*cf.* 4.1).

Le corpus [STRO] est composé de 18 combinaisons /pVCV/ avec V = [a i u], C = [b d m n ʁ l] prononcés par un locuteur français PB (Voyelle-Consonne-Voyelle). Il a été obtenu par la méthode d'échantillonnage synchronisé décrite par Masaki *et al.* (1999), dans laquelle le locuteur répète 128 fois chaque séquence en synchronie avec un bip délivré par l'imageur (voir Fig. 1a pour un exemple d'image). Le locuteur a été enregistré en 2002 dans les laboratoires ATR Human Information Processing Research Laboratories (Kyoto, Japan) (imageur : Marconi Eclipse 1,5 T, 468 images médio-sagittales, 28,9 images/sec reconstruites, résolution de 1 mm/px, champ de vue 256×256 mm², épaisseur de coupe 5 mm, profondeur d'image 8 bits/pixel, technique d'écho de gradient, temps de répétition (TR) 900 ms, temps d'écho (TE) 3 ms, angle de bascule (Flip) 30°).

Le corpus [VCV30] comprend 5 répétitions de CV avec C = [p t k] et V = [a i u] et de [fa sa ʁa ʒu zi ma la bao] prononcées par une locutrice allemande NB au Biomedizinische NMR Forschungs GmbH Göttingen. Les images fournies par une séquence d'écho de gradient radial hautement sous-échantillonné (3T Siemens Prisma Fit MRI System, 3225 images, 30 im/sec, 1,41 mm/px, 192×192 mm², épaisseur de coupe 8 mm, 12 bits/px, TR 1,96 ms, TE 1,28 ms, Flip 5°, nombre de projections radiales par image reconstruite 17 (NP)) sont reconstruites conjointement avec les profils de sensibilité des antennes en résolvant un problème inverse non linéaire suivant la méthode décrite dans Uecker *et al.* (2010).



(a) (b) (c) (d)

FIGURE 1 : (a) image [STRO] avec un exemple de cadres (cadre global jaune) ; (b, haut) zoom sur le cadre blanc de (a) avec contour de langue tracé et segment de normale au contour sous-tendant le profil d'intensités ; (b, bas) zoom avec contour prédit et illustration des segments de normale pour l'exploration des profils; (c, haut) exemple de profils moyens ; (c, bas) profils correspondant à (b, bas) ; (d) image [VCV55] avec exemple de contours prédits.

Le corpus [VCV55] comprend les mêmes répétitions de CV produites par la même locutrice. Le taux d'acquisition est de 55 images/sec (3200 images, TR 2,0 ms, NP 9). Notons qu'aucune différence visible n'apparaît entre les images de VCV30 et VCV55 (à l'exception de la mâchoire dans les transitions rapides). Les images ont été ensuite sur-échantillonnées à 0,71 mm/px (voir exemple à la Fig. 1d).

Notons que toutes les images ont été débruitées avant tout autre traitement à l'aide d'ondelettes de Daubechies de type 1. Elles sont ensuite recalées par rapport au palais grâce à la comparaison à un motif standard précédemment explicitée. Elles sont finalement recadrées comme indiqué en 4.2.

5.2 Résultats d'évaluation

Nous avons testé plusieurs combinaisons des méthodes décrites ci-dessus. Tous les résultats ont été obtenus par une méthode de validation croisée (*leave-one-out cross validation, cf.* Arlot & Celisse (2010)) qui calcule l'erreur d'estimation pour chaque élément test en utilisant les $n_{cl} - 1$ autres éléments pour établir les modèles. Cette procédure a été appliquée sur les ensembles d'apprentissage déterminés par classification hiérarchique pour chaque corpus. Nous avons utilisé deux métriques pour mesurer les différences entre contours : la RMS des distances points à points des contours, et la RMS des distances des points de contours prédits aux contours tracés.

Deux types de modèles linéaires de contours déformables ont été testés : (1) un modèle *global* M_{gbl} qui représente l'ensemble des coordonnées des contours avec un seul jeu de $n_{int\_gbl} = 12$ composantes donnant une erreur RMS de distance point à point moyennée sur les organes entre 0,86 et 1,06 mm (et de 0,60 – 0,74 mm pour les distances points à contour) suivant les corpus, et (2) un ensemble de modèles d'organes *locaux* M_{orgs}, donnant des erreurs de 0,23 – 0,26 mm (0,20 – 0,22) avec n_{cnt_orgs} totalisant entre 44 et 53 composantes (partiellement corrélées entre organes) qui expliquent chacune au moins 0.1% de la variance par organe (et au total au moins 99,5% de la variance par organe).

De manière analogue, nous avons testé deux types de modèles MLR : un modèle *global* MLR_{gbl} qui prédit les coordonnées des contours par l'intermédiaire de M_{gbl} à partir des $n_{int\_gbl} = 25$ composantes représentant environ 95% de la variance des intensités de l'image cadrée sur l'ensemble du conduit vocal avec une erreur de 1,48 – 1,56 mm (0,85 – 0,91) suivant les corpus, et des modèles *locaux* MLR_{orgs} qui prédisent les contours de chaque organe par l'intermédiaire des M_{orgs} à partir des n_{int_orgs} (entre 3 et 40) composantes des images cadrées sur chaque organe donnant une erreur de 1,23 – 1,41 mm (0,71 – 0,82).

Nous avons également testé l'amélioration apportée à la prédiction des modèles MLR par notre méthode ASMM. L'application de l'ASMM à chacun des contours prédits par le modèle *global* MLR_{gbl} permet de réduire les erreurs à 1,34 – 1,37 mm (0,64 – 0,67), tandis que les erreurs des modèles M_{orgs} sont ramenées à 1,14 – 1,31 mm (0,59 – 0,65). Pour les organes séparés, les erreurs varient de 0,59 à 2,04 mm (0.33 – 1.03), avec une gamme de 1,73 à 2,04 mm (0.91 – 1.03) pour la langue qui a l'erreur la plus grande. Un exemple de contours obtenus est donné à la Fig. 1d. Notons que les estimations des erreurs RMS sont globales et cachent des disparités entre phonèmes, positions sur les organes, et aussi corpus et sujets ; les erreurs les plus importantes se retrouvent sur les extrémités de la langue par exemple (jusqu'à 2,91 mm point à point et 2,43 mm point à contour).

Notons que la méthode des démons a donné des erreurs de 1,90 – 2,09 (1,02 – 1,12), plus élevées que les méthodes MLR avec ASMM, et qui ne sont pas suffisamment réduites par les ASMM.

Nous avons également testé plusieurs méthodes pour les organes rigides (mâchoire et hyoïde). La méthode MLR suivie d'ASMM donne une erreur de 0,72 – 1,15 mm (0,42 – 0,59) pour la mâchoire, et de 0,93 – 1,53 mm (0,59 – 1,07) pour l'hyoïde. Ces résultats sont assez similaires à ceux obtenus pour les organes déformables. Les résultats les meilleurs sont obtenus par des méthodes différentes en fonction des corpus, souvent le démon suivi d'un ASM, mais ne sont pas nettement meilleurs que pour la méthode MLR + ASMM : pour la mâchoire les erreurs sont de 0,62 – 0,86 mm (0,40 – 0,45), et pour l'hyoïde de 0,93 – 1,37 (0,59 – 0,83). Les méthodes basées sur la correspondance à un motif standard sont moins performantes.

6 Conclusion et perspectives

Nous avons développé une méthode de prédiction des contours individuels médiosagittaux des principaux organes orofaciaux impliqués dans la parole et la déglutition (le palais dur, la langue, les lèvres supérieure et inférieure, le velum, la paroi arrière pharyngée, l'épiglotte, ainsi que la mâchoire et l'os hyoïde) à partir d'images IRM dynamiques. Cette méthode est basée sur l'apprentissage de modèles d'apparence et de modèles de forme, ainsi que d'un modèle multilinéaire dont les résultats sont ensuite corrigés par un ASM modifié utilisant les informations locales de profils d'intensité de pixels sur les normales aux contours. Cette nouvelle méthode donne des erreurs moyennes points à contour entre 0,57 et 0,70 mm tous organes confondus. Ces performances, nettement meilleures que celles décrites dans les articles qui donnent des évaluations chiffrées, sont atteintes au prix d'un tracé manuel de tous les contours pour un corpus d'apprentissage d'une soixantaine d'images. Cet inconvénient est cependant minime si l'on souhaite traiter des corpus de centaines de milliers d'images pour le même locuteur. Nous notons aussi que les résultats sont très sensiblement semblables pour deux méthodes d'IRM aussi différentes que la méthode d'échantillonnage synchronisé qui oblige le sujet à répéter 128 fois le segment de phrase et la méthode d'écho de gradient radial hautement sous-échantillonné suivi de reconstruction par inversion non linéaire qui permet des acquisitions en temps-réel jusqu'à 55 images/sec. ou plus.

Cette nouvelle méthode ouvre des perspectives très intéressantes. Nous allons tout d'abord la tester pour des tâches de déglutition qui risquent de s'avérer plus difficiles à traiter, parce que les organes sont souvent en contact entre eux ou avec les aliments, et donc les contrastes plus faibles. Les contours obtenus pour la parole permettront d'établir des modèles articulatoires plus élaborés et d'analyser plus finement la variabilité et la coarticulation en parole. Les quantités de données possibles permettront également d'établir par apprentissage automatique des cartes d'association entre diverses modalités pour un même locuteur (p. ex. articulation, son) ou entre locuteurs.

Remerciements

Ce travail a bénéficié du support de l'ANR par les projets ANR-13-TECS-0011-06 «e-SwallHome» et ANR-11-INBS-0006 «Infrastructure d'avenir en Biologie Santé».

Références

ARLOT, S. & CELISSE, A. (2010). A survey of cross-validation procedures for model selection. 40-79.

BRESCH, E. & NARAYANAN, S. (2009). Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. *IEEE Transactions on Medical Imaging, 28*(3), 323-338.

COOTES, T.F., TAYLOR, C.J., COOPER, D.H. & GRAHAM, J. (1995). Active shape models - Their training and application. *Computer Vision and Image Understanding, 61*(1), 38-59.

ERYILDIRIM, A. & BERGER, M.-O. (2011). A guided approach for automatic segmentation and modeling of the vocal tract in MRI images. In *19th European Signal Processing Conference (EUSIPCO 2011)* pp. 61-65. Barcelona, Spain.

KIM, J., KUMAR, N., LEE, S. & NARAYANAN, S.S. (2014). Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data. In *10th International Seminar on Speech Production, ISSP10* (S. Fuchs, M. Grice, A. Hermes, L. Lancia & D. Mücke, Eds.), pp. 222-225. Cologne, Germany.

KROON, D.-J. & SLUMP, C.H. (2009). MRI Modality transformation in demon registration, *IEEE International Symposium on Biomedical Imaging, ISBI '09* (pp. 963-966). Boston, MA: IEEE Signal Processing Society.

MASAKI, S., TIEDE, M.K., HONDA, K., SHIMADA, Y., FUJIMOTO, I., NAKAMURA, Y. & NINOMIYA, N. (1999). MRI-based speech production study using a synchronized sampling method. *Journal of the Acoustical Society of Japan (English), 20*(5), 375-379.

NIEBERGALL, A., ZHANG, S., KUNAY, E., KEYDANA, G., JOB, M., UECKER, M. & FRAHM, J. (2013). Real-Time MRI of speaking at a resolution of 33 ms: Undersampled radial FLASH with nonlinear inverse reconstruction. *Magnetic Resonance in Medicine, 69*, 477-485.

OLTHOFF, A., ZHANG, S., SCHWEIZER, R. & FRAHM, J. (2014). On the physiology of normal swallowing as revealed by magnetic resonance imaging in real time. *Gastroenterology Research and Practice, 2014*, 10.

PROCTOR, M.I., BONE, D., KATSAMANIS, A. & NARAYANAN, S.S. (2010). Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis. In *Interspeech 2010 (11th Annual Conference of the International Speech Communication Association)* (T. Kobayashi, K. Hirose & S. Nakamura, Eds.), pp. 1576-1579. Makuhari, Japan.

SERRURIER, A. & BADIN, P. (2008). A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data. *Journal of the Acoustical Society of America, 123*(4), 2335-2355.

SILVA, S. & TEIXEIRA, A. (2015). Unsupervised segmentation of the vocal tract from real-time MRI sequences. *Computer Speech & Language, 33*(1), 25-46.

UECKER, M., ZHANG, S., VOIT, D., KARAUS, A., MERBOLDT, K.-D. & FRAHM, J. (2010). Real-time magnetic resonance imaging at a resolution of 20 ms. *NMR in Biomedicine 23*, 986–994.

Sur les traces acoustiques de /ʃ/ et /ç/ en allemand L2

Jane Wottawa [1] Martine Adda-Decker[1]
(1) 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

Les apprenants français de l'allemand ont des difficultés à produire la fricative palatale sourde allemande /ç/ (*Ich*-Laut) et ont tendance à la remplacer par la fricative post-alvéolaire /ʃ/. Nous nous demandons si avec des mesures acoustiques ces imprécisions de production peuvent être quantifiées d'une manière plus objective. Deux mesures acoustiques ont été examinées afin de distinguer au mieux /ʃ/ et /ç/ dans un contexte VC en position finale de mot dans des productions de locuteurs germanophones natifs. Elles servent ensuite à quantifier les difficultés de production des apprenants français. 285 *tokens* de 20 locuteurs natifs et 20 locuteurs L2 ont été analysés. Les mesures appliquées sont le centre de gravité spectral et des rapports d'intensité par bande de fréquence. Sur les productions de locuteurs natifs, les résultats montrent que la mesure la plus fiable pour distinguer acoustiquement /ʃ/ et /ç/ est le ratio d'intensité entre fréquences hautes (4-7 kHz) et basses (1-4 kHz). Les mesures confirment également les difficultés de production des locuteurs natifs français.

ABSTRACT

Acoustic tracing of /ʃ/ and /ç/ in German L2.

French learners of German often replace the unvoiced palatal fricative /ç/ (*Ich*-Laut) by the post-alveolar fricative /ʃ/. We are interested in whether acoustic measurements can help quantifying these production imprecisions in an objective way. We examined two acoustic measures to distinguish between /ʃ/ and /ç/ in a word-final VC position. A total of 285 tokens of 20 native and 20 non-native speakers were analyzed. The applied measures are the spectral centre of gravity and intensity per frequency band. We introduced an intensity ratio between high (4-7kHz) and low (1-4 kHz) frequency bands. On native speech, results show that the most reliable measure to distinguish between /ʃ/ and /ç/ is the intensity ratio between high and low frequency bands. The measurements also confirm the production difficulties of the French native speakers.

MOTS-CLÉS : fricatives, mesures acoustiques, intensité, allemand, français, production en L2.

KEYWORDS: fricatives, acoustic measures, intensity, German, French, L2 production.

1 Introduction

Apprendre une langue étrangère ne se limite pas à l'apprentissage des règles de grammaire et du vocabulaire. La prononciation est aussi importante pour s'intégrer dans une communauté linguistique. Dans notre recherche, nous nous intéressons à la prononciation de l'allemand par des apprenants francophones vivant en France métropolitaine. L'allemand et le français n'ont pas le même inventaire phonétique : l'allemand standard ne connaît pas de voyelles nasales, le français ne compte pas /h/, /ŋ/ et /ç/ (*Ich*-Laut) parmi ses consonnes. Les apprenants des deux langues ont souvent des difficultés à produire des sons qui sont absents de l'inventaire phonétique de leur langue maternelle même s'ils sont capables de les apprendre grâce à un entraînement approprié (Flege *et al.*, 1995). Afin de fournir un *feedback* de prononciation fiable et propre à chaque apprenant, nous voulons trouver des mesures acoustiques qui nous permettent de comparer leur production orale à l'allemand standard. C'est pourquoi, nous nous intéressons dans cet article à la production du /ç/ dans des mots allemands par des germanophones (GG) et des francophones (FG). Les FG ont tendance à produire la fricative post-alvéolaire /ʃ/ existant en français au lieu de /ç/ absent de l'inventaire phonémique du français ou à généraliser la production de /ç/ une fois que celle-ci est bien acquise.

1.1 Les fricatives allemandes /ʃ/ et /ç/

L'allemand possède plusieurs fricatives qui ne font pas partie de l'inventaire consonantique du français. En plus du /h/ qui pose des problèmes bien connus aux natifs français (Zimmerer & Trouvain, 2015; Wottawa *et al.*, 2015), il y a deux fricatives sourdes, une palatale (*Ich*-Laut) et une vélaire (*Ach*-Laut), qui sont considérées comme allophones du même phonème, dans la mesure où elles apparaissent en distribution complémentaire dans le lexique : le [ç] se trouve après les voyelles antérieures (et consonnes) et le [x] après les voyelles postérieures et le /a/ (Kohler, 1990), en général en fin de mot (*Buch* [bux], livre) ou en fin de morphème (*riech-en* [ʁiç-ən], sentir). Nous nous intéressons ici à la fricative palatale sourde qui est souvent remplacée par les locuteurs natifs français par son plus proche voisin, la post-alvéolaire /ʃ/. Dans la langue allemande on trouve peu de paires minimales (*fischte* [fɪʃtə] pêcher (participe) - *Fichte* [fɪçtə], épicéa ; *misch* [mɪʃ] melanger (vocatif) - *mich* [mɪç] moi) et les germanophones natifs eux-mêmes sont parfois amenés à confondre les deux sons.

La fricative post-alvéolaire /ʃ/ peut se trouver soit au début soit à la fin d'une syllabe allemande, par exemple : *schnell* [ʃnɛl] (rapide) et *Fisch* [fɪʃ] (poisson) et dans le suffixe *-isch*. La fricative palatale sourde /ç/ forme souvent un cluster consonantique avec la plosive /t/ à la fin des mots monosyllabiques : *Licht* [lɪçt] (lumière), *echt* [ɛçt]. Dans la morphologie dérivationnelle, les suffixes *-chen* et *-(l)ich* contiennent également la palatale sourde /ç/. Nous nous concentrons ici sur la réalisation des suffixes *-isch* [ɪʃ] et *-(l)ich* [(l)ɪç] par des germanophones natifs et non-natifs.

Dans figure 1 les mots *solidarisch* et *freundlich*, illustrés par des spectrogrammes, ont été extraits à partir du même fichier son. Les différences observées ne résultent alors pas de conditions d'enregistrement différentes. Le dernier segment, au niveau phonémique, pour les deux mots correspond aux fricatives que nous essayons distinguer par des mesures acoustiques. /ʃ/ se distingue du /ç/ par une intensité plus importante qui se traduit par une portion de bruit plus noircie dans le spectrogramme. Mais nous observons également dans le spectrogramme que les deux fricatives /ʃ/ et /ç/ semblent se situer sur la même plage de fréquence.

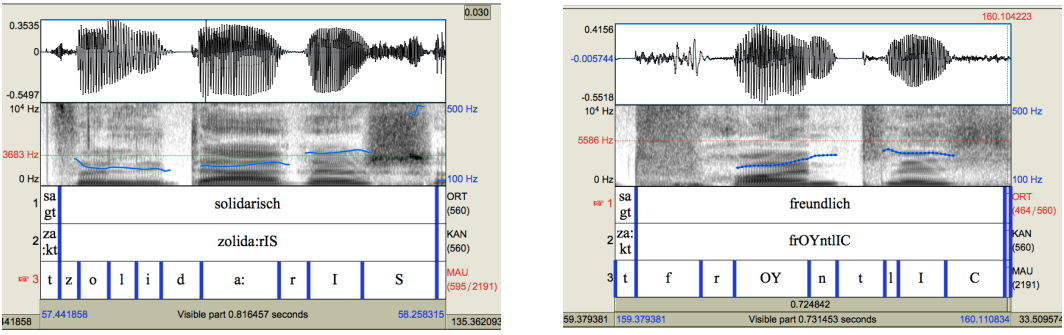


FIGURE 1 – Spectrogrammes de mots en -[ɪʃ] et -[ɪç] - GG, féminin

A la suite, le choix du matériel acoustique est expliqué, suivie d'une présentation des mesures acoustiques du centre de la gravité et de l'intensité par bande de fréquence avec leurs résultats respectifs après leur application aux productions des locuteurs germanophones natifs et non-natifs.

2 Choix du matériel acoustique

Nous avons extrait les mots qui portent les suffixes -/ɪʃ/ ou -/ɪç/ tels que *solidarisch* [zoliˈda ːrɪʃ] (solidaire) ou *freundlich* [ˈfr͡ɔɪntlɪç] (aimable) du *French Learners Audio Corpus of German Speech* (FLACGS) enregistré au Laboratoire de Phonétique et de Phonologie, Paris 3 (LPP) entre 2014 et 2015, . Ainsi un ensemble de 285 *tokens* de 40 locuteurs dont 20 natifs et 20 non-natifs a été analysé. Cette étude s'intéresse uniquement aux mots produits dans la tâche de répétition. Ce matériel a l'avantage de présenter tous les mots cibles au milieu d'une des phrases cadre : *Er sagt ... klar und deutlich* ou *Ich sage ... klar und deutlich*. Les différences acoustiques que nous observons entre les mots en -/ɪʃ/ et -/ɪç/ ne sont alors pas liées à la position du mot dans la phrase. Tous les mots cibles produits dans la tâche de répétition portent l'accent de phrase même si les suffixes eux-mêmes ne sont pas accentués. Les fricatives /ʃ/ et /ç/ sont précédées par la voyelle /ɪ/ et suivies d'une pause et de la plosive /k/.

Locuteurs	GG				FG			
	femmes		hommes		femmes		hommes	
Canonique	[ʃ]	[ç]	[ʃ]	[ç]	[ʃ]	[ç]	[ʃ]	[ç]
[ʃ]	100,0 %	0,0 %	96,7 %	0,0 %	73,3 %	40 %	53,3 %	22,5 %
[ç]	0,0 %	87,5 %	0,0 %	87,5 %	16,7 %	45 %	33,3 %	72,5 %
Ambiguë	0,0 %	12,5 %	3,3 %	12,5 %	10 %	15 %	13,3 %	5 %

TABLE 1 – Perception des suffixes /ɪʃ/ et /ɪç/ chez les GG et les FG

Les 285 *tokens* (les mots entiers) ont été jugés par un locuteur germanophone natif selon la prononciation du suffixe. Dans Table 1 les résultats de ce jugement perceptif sont résumés. Le tableau dégage trois catégories perceptives pour les fricatives canoniques /ʃ/ et /ç/. La troisième catégorie qui s'ajoute est une fricative qui pourrait se situer entre les deux fricatives canoniques et n'est pas parfaitement identifiable comme /ʃ/ ou /ç/. Chez les germanophones natifs, cette catégorie est plutôt une variante

pour la fricative palatale sourde /ç/ qui est produite autant par les femmes que les hommes. Chez les non-natifs, nous observons des confusions entre /ʃ/ et /ç/ qui sont absentes dans le groupe de natifs pour la tâche de répétition. Des substitutions de /ç/ par /ʃ/ ont été observées par d'autres chercheurs qui s'intéressent à l'apprentissage de l'allemand par des francophones (Jouvet *et al.*, 2015). Le fait que /ʃ/ à son tour puisse être remplacé par /ç/ n'est par mentionné par Jouvet *et al.* (2015). Les locuteurs francophones qui ont effectué cette substitution n'ont souvent pas substitué les fricatives /ç/ canoniques par /ʃ/. Il est alors possible que les substitutions de /ʃ/ par /ç/ indiquent une hypercorrection de la part des apprenants.

3 Mesures acoustiques

Les mesures acoustiques menées sur le sous-ensemble du corpus présenté dans section 2 ont pour objectif de définir si un locuteur non-natif (ou même natif) produit la fricative canonique dans les différents mots.

Les mesures acoustiques ont été menées avec le logiciel Praat (Boersma & Weenink, 2016). Dans ce cadre, les fichiers sons ont été transcrits manuellement selon l'orthographe allemande, puis la transcription a été alignée automatiquement avec le service web MAUS (*Munich Automatic Segmentation*) (Schiel, 1999; Kisler *et al.*, 2012). A la fin, une vérification manuelle a été effectuée sur des mots choisis.

Nous allons à présent aborder les mesures du centre de gravité et des rapports d'intensité par bande de fréquence d'une part avec la voyelle précédant la fricative et d'autre part entre les fréquences hautes et basses de la fricative.

3.1 Centre de gravité

Le centre de gravité correspond à la moyenne fréquentielle pondérée par son amplitude. Cette mesure est utile si l'on veut distinguer les différentes fricatives tel que cela a été proposé par différentes équipes de recherche (Żygis & Padgett, 2010; Kemp, 2011; Benamrane, 2013). Afin de savoir comment le centre de gravité évolue sur l'ensemble de la fricative, nous avons extrait le centre de gravité du début, du milieu et de la fin de la fricative avec Praat. Les valeurs ont ensuite été moyennées sur la population entière.

3.2 Bandes de fréquences

La figure 1 montre que /ʃ/ a une intensité plus élevée que /ç/. Les mesures du centre de gravité qui sont présentées dans section 4.1 suggèrent qu'une analyse complémentaire est nécessaire pour distinguer les deux fricatives allemandes /ʃ/ et /ç/. Dans le spectrogramme, nous observons que l'intensité n'est pas distribuée à part égale sur toutes les bandes de fréquences pour les fricatives /ʃ/ et /ç/. Nous proposons alors une analyse par bande de fréquence. Avec Praat, l'intensité de la fricative est extraite entre 1-3kHz, 1-4kHz, 3-6kHz et 4-7kHz. Nous observons dans un premier temps les valeurs brutes par bande de fréquence, puis nous calculons le ratio de l'intensité par bande de fréquence entre la fricative et la voyelle précédente enfin, nous proposons un ratio entre bandes de fréquences hautes et basses.

4 Résultats

4.1 Centre de gravité

Les analyses du centre de gravité ne montrent pas de différence entre les deux fricatives [ʃ] et [ç] dans les productions des germanophones natifs. En effet, comme nous pouvons l'observer dans figure 2, la seule différence entre [ʃ] et [ç] est la largeur de la plage spectrale occupée par le centre de gravité de la fricative : chez les natifs [ʃ] occupe une plage qui se situe entre 2898 et 4038 Hz si on regarde les valeurs du deuxième et troisième quartile tandis que [ç] occupe une plage qui se situe entre 2379 et 5283 Hz. Chez les non-natifs, la plage occupée par le centre de gravité de la fricative [ʃ] est plus large que chez les germanophones natifs. Elle se situe entre 2337 et 4066 Hz. Nous observons aussi plus de variabilité entre le début, le milieu et la fin de la fricative chez les non-natifs. Chez les non-natifs, la plage du centre de gravité se restreint au fur et à mesure tandis qu'elle reste constante chez les germanophones natifs. Concernant le centre de gravité de la fricative [ç], chez les deux groupes de locuteurs, la plage occupée est large mais chez les natifs elle se situe dans une bande de fréquence plus élevée que chez les non-natifs : pour les valeurs les plus basses, il y a une différence de plus que 600 Hz et pour les valeurs les plus élevées la différence excède les 1000 Hz. Dans les productions des non-natifs, nous observons les valeurs maximales du centre de gravité des [ç] qui correspondent à ceux de leur [ʃ] tandis que les germanophones natifs ont tendance à produire des [ç] dont les valeurs les plus élevées du centre de gravité dépassent nettement celles observées pour le [ʃ] des locuteurs natifs. Chez les deux groupes de locuteurs, la figure 2 montre une variabilité du centre de gravité pour le début, le milieu et la fin de [ç] dont le sommet se trouve sur la portion du milieu. Chez les natifs, le sommet est plus prononcé que chez les non-natifs.

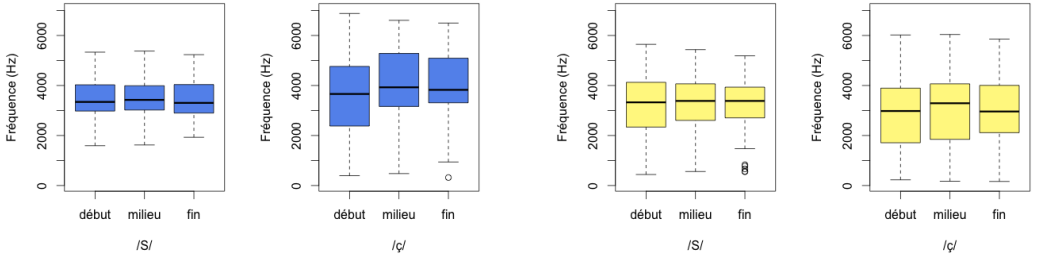


FIGURE 2 – Centre de gravité pour les trois portions de fricative - GG à gauche et FG à droite

Malgré les différences spectrales du centre de gravité entre [ʃ] et [ç], cette mesure acoustique, à elle seule, n'est pas fiable pour distinguer les deux fricatives. Les plages de fréquences occupées par les centres de gravité des deux consonnes se chevauchent. Une distinction tranchée est alors impossible aussi bien pour les natifs que pour les non-natifs. L'inventoire consonantique gaélique contient également les fricatives /ʃ/ et /ç/ (Gordon *et al.*, 2002). Les chercheurs rapportent également des centres de gravité très proches pour la moyenne de groupe. Le centre de gravité ne semble pas être une mesure acoustique suffisamment précise pour distinguer les fricatives palatales sourdes et les fricatives post-alvéolaires.

4.2 Bandes de fréquences

4.2.1 Valeurs absolues

Comme pour le centre de gravité, les mesures d'intensité par bande de fréquence sont prises à partir du spectre. La figure 3 montre l'intensité moyenne par bande de fréquence des fricatives [ʃ] et [ç] pour les locuteurs natifs et non-natifs. Les bandes de fréquences sont indiquées sur l'abscisse en kHz.

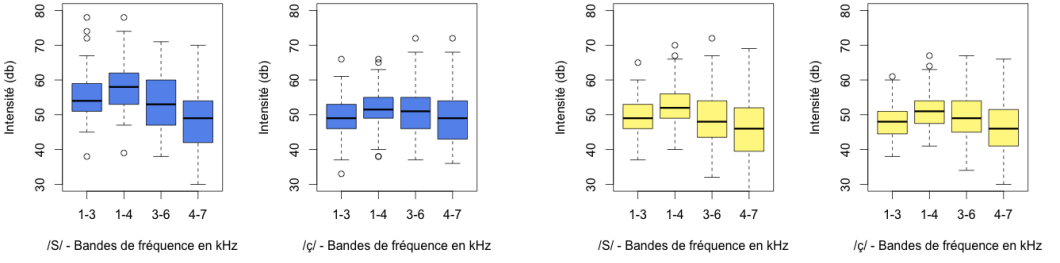


FIGURE 3 – Intensité par bande de fréquence - GG à gauche et FG à droite

Les moyennes des valeurs absolues présentées en figure 3 ne donnent pas de séparation précise entre [ʃ] et [ç], ni pour les natifs ni pour les non-natifs. Pour les deux groupes de locuteurs, nous observons moins de chevauchement pour la bande de fréquence qui se situe entre 1 et 4 kHz. C'est sur cette bande de fréquence que nous calculons le ratio de l'intensité entre la voyelle [ɪ] et la fricative suivante [ʃ] ou [ç].

4.2.2 Ratio entre [ʃ]/[ɪ] ou [ç]/[ɪ]

La table 2 récapitule le taux d'identification de [ʃ] et [ç] dans les suffixes /ɪʃ/ et /ɪç/ chez les natifs et les non-natifs. La classification semble plutôt solide si l'on se base sur la représentation canonique des fricatives. En revanche, les valeurs présentées dans la table 2 ne correspondent pas aux valeurs de perception que nous avons pu observer dans la table 1. Seul, le ratio établi entre la fricative et la voyelle précédente ne nous donne pas des résultats fiables pour décider si les locuteurs produisent [ʃ] ou [ç]. La fiabilité réduite de la mesure est étroitement liée à la grande variabilité qui existe entre les locuteurs (même natifs) dans leur productions de [ɪʃ] et [ɪç].

Locuteurs	GG				FG			
	femmes		hommes		femmes		hommes	
Canonique	[ʃ]	[ç]	[ʃ]	[ç]	[ʃ]	[ç]	[ʃ]	[ç]
[ʃ]	92,6 %	2,8 %	97,0 %	29,5 %	69,7 %	51,1 %	69,2 %	47,1 %
[ç]	7,4 %	97,2 %	3,0 %	70,5 %	30,3 %	48,9 %	30,8 %	52,9 %

TABLE 2 – Reconnaissance de [ʃ] et [ç] avec le ratio C/V, bande : 1-4kHz, seuil = 1,03

4.2.3 Ratio entre les bandes de fréquences hautes et basses

Afin de nous assurer que les résultats peu stables de la section 4.2.2 ne sont pas liés à la qualité vocalique, nous présentons maintenant le ratio entre l'intensité des bandes de fréquences hautes et basses pour les fricatives [ʃ] et [ç]. Nous divisons les mesures effectuées en deux et analysons d'abord le classement par les fricatives canoniques et ensuite le classement selon le phonème perçu.

Figure 4 illustre les différences qui existent entre le classement des phonèmes canoniques et les phonèmes perçus. Les différences sont surtout visibles pour les natifs. A la fois pour les phonèmes canoniques et les phonèmes perçus, nous observons de grands écart-types qui suggèrent qu'un classement précis entre les fricatives allemandes [ʃ] et [ç] est compliqué. Encore une fois, la variabilité entre les locuteurs est trop importante pour décider d'une manière fiable par des mesures acoustiques si la fricative produite est [ʃ] ou [ç].

D'après les résultats du ratio entre l'intensité des fréquences hautes et basses, il semble exister une troisième catégorie de fricative au moins pour les natifs. Dans nos données, les natifs remplacent 12,5% de leur productions de [ç] par une autre variante (cf. table 1). La partie droite de la figure 4 suggère que cette variante ne se situe pas entre les fricatives [ʃ] et [ç] au moins pour les natifs mais qu'elle est plutôt associée à la catégorie de [ç].

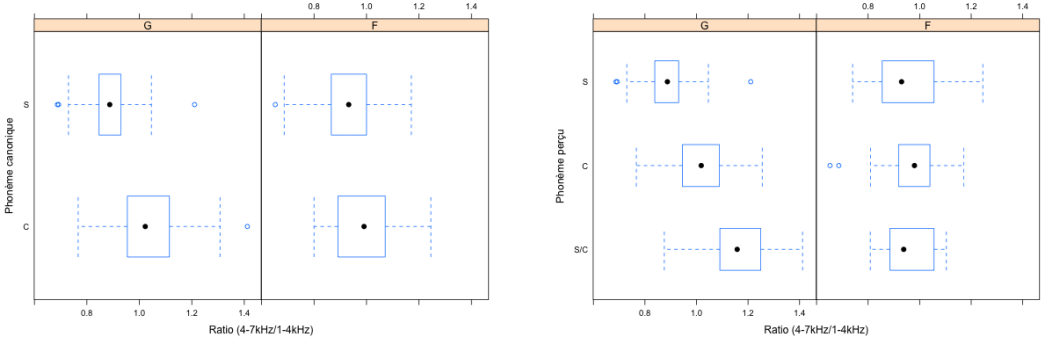


FIGURE 4 – Ratio entre les fréquences hautes et basses - canonique à droite, perçu à gauche

Dans la table 3 sont résumés les résultats pour le classement des fricatives canoniques et perçus à un seuil de 0,96 pour le ratio entre l'intensité des fréquences hautes et basses. Le tableau montre que les valeurs sont plus précises pour les phonèmes perçus la mesure ne suffit pas à elle seule pour que nous puissions distinguer [ʃ] et [ç] d'une manière précise. Pour les non-natifs, la classification des phonèmes perçus n'est pas meilleure que celle des phonèmes canoniques. Dans la figure 4, nous pouvons observer que les trois catégories de perception chez les non-natifs ne sont pas bien distinctes pour le ratio d'intensité entre fréquences hautes et basses. Il semble que la troisième catégorie perçue chez les non-natifs soit en effet une variante de fricative qui se situe entre [ʃ] et [ç].

La table 3 suggère également que les groupes de locuteurs pour l'analyse de l'intensité ne doivent pas seulement être séparés en locuteurs germanophones natifs et non-natifs mais aussi en hommes et femmes.

Locuteurs	GG					
	femmes			hommes		
Canonique	[ʃ]	[ç]	A	[ʃ]	[ç]	A
Classification [ʃ] (canonique)	75,0 %	0,0 %	*nd*	81,3 %	50 %	*nd*
Classification [ç] (canonique)	25,0 %	100 %	*nd*	18,7 %	50 %	*nd*
Classification [ʃ],[ç], A (perception)	87,1 %	100 %	100 %	86,2 %	45,7 %	66,7 %

Locuteurs	FG					
	femmes			hommes		
Canonique	[ʃ]	[ç]	A	[ʃ]	[ç]	A
Classification [ʃ] (canonique)	39,4 %	26,1 %	*nd*	74,1 %	54,9 %	*nd*
Classification [ç] (canonique)	60,6 %	73,9 %	*nd*	25,9 %	45,1 %	*nd*
Classification [ʃ],[ç], A (perception)	40,5 %	75,0 %	75,0 %	79,2 %	42,1 %	0,0 %

TABLE 3 – Reconnaissance par le ratio entre les fréquences 4-7 kHz et 1-4 kHz, seuil = 0,96

5 Discussion et conclusion

Afin de distinguer les productions de [ʃ] et [ç] chez les natifs et non-natifs, nous avons effectué des mesures du centre de gravité et de l'intensité par bande de fréquences.

Concernant les locuteurs natifs, [ʃ] et [ç] sont le mieux séparés par les mesures d'intensité prenant en compte la voyelle précédente dans la bande de fréquence de 1-4 kHz. En effet, l'intensité de [ʃ] est maximale dans cette plage de fréquence (cf. figure 3). La mesure livre des résultats très robustes pour les femmes. Chez les hommes, [ç] est parfois confondu avec [ʃ] (29,5%). Cela peut être lié à différents facteurs. D'une part, il semble que les hommes produisent les fricatives palatales avec une intensité plus élevée que les femmes, d'autre part, les occurrences de [ç] qui étaient ambiguës en perception (cf. table 1) ont été classifiées en tant que [ʃ] chez les hommes. Avec la même approche, la classification des productions non-natives de [ç] se situe seulement autour de 50%. Ce résultat suggère que la mesure contextuelle n'aide pas à trancher entre [ʃ] et [ç] chez les non-natifs.

Pour les non-natifs, les meilleurs résultats pour la distinction de [ʃ] et [ç] est le ratio entre les fréquences 4-7 kHz (hautes) et 1-4 kHz (basses) (cf. table 3). Nous observons une différence entre hommes et femmes.Avec cette méthode, le [ç] est mieux classifié pour les femmes que pour les hommes. Concernant le [ʃ], les productions des hommes sont mieux classifiées que celles des femmes.

Distinguer les productions de [ʃ] et [ç] avec des mesures acoustiques est possible pour les natifs même si la variabilité inter-locuteurs est importante. En revanche, même si la différence entre [ʃ] et [ç] des non-natifs est perceptible, les mesures acoustiques qui ont été choisies ne suffissent pas à juger d'une manière fiable si les non-natifs produisent [ʃ] ou [ç]. Les résultats suggèrent qu'une séparation selon le genre peut améliorer la classification globale.

Certaines autres analyses acoustiques sont prévues afin de distinguer [ʃ] et [ç] : traditionnellement la transition formantique est mesurée (Delattre *et al.*, 1962; Żygis & Padgett, 2010; Benamrane, 2013), un calcul de variabilité du centre de gravité au fil de la fricative ainsi que la mesure de l'amplitude dynamique (Shadle & Mair, 1996). Une combinaison de différentes mesures acoustiques (contextuelles et non-contextuelles) devrait permettre une meilleure classification des deux fricatives.

Remerciements

Ce travail a été soutenu par le programme Investissements d'Avenir - Labex EFL program (ANR-10-LABX-0083) et par l'Université Sorbonne Nouvelle et l'ED 268 par un contrat doctoral.

Références

BENAMRANE A. (2013). *Acoustic study of fricatives in standard Arabie (Algerian speakers)*. Theses, Université de Strasbourg.

BOERSMA P. & WEENINK D. (2016). *Praat : doing phonetics by computer [Computer program]*. Version 6.0.15, retrieved 23 March 2016.

DELATTRE P. C., BERMAN A. & COOPER F. S. (1962). Formant transitions and loci as acoustic correlates of place of articulation in american fricatives. *Studia Linguistica*, **16**(1-2), 104–122.

FLEGE J. E., TAKAGI N. & MANN V. (1995). Japanese adults can learn to produce english /i/ and /l/ accurately. *Language and Speech*, **38**(1), 25–55.

GORDON M., BARTHMAIER P. & SANDS K. (2002). A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association*, **32**(02), 141–174.

JOUVET D., BONNEAU A., TROUVAIN J., ZIMMERER F., LAPRIE Y. & MÖBIUS B. (2015). Analysis of phone confusion matrices in a manually annotated french-german learner corpus. In *Workshop on Speech and Language Technology in Education*.

KEMP R. L. (2011). *The Perception of German Dorsal Fricatives by Native Speakers of English*. PhD thesis, Master thesis, University of Georgia. Cerca con Google.

KISLER T., SCHIEL F. & SLOETJES H. (2012). Signal processing via web services : the use case webmaus. In *Digital Humanities Conference 2012*.

KOHLER K. (1990). German. *Jurnal of the International Phonetic Association*, **20**(01), 48–50.

SCHIEL F. (1999). Automatic phonetic transcription of non-prompted speech. In *Proc. Int. Cong. Phon. Sci*, p. 607–610.

SHADLE C. H. & MAIR S. J. (1996). Quantifying spectral characteristics of fricatives. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, p. 1521–1524 : IEEE.

WOTTAWA J., ADDA-DECKER M. & ISEL F. (2015). Segmental difficulties in french learners of german. In *Proceedings of the International Symposium on Monolingual and Bilingual Speech 2015*.

ZIMMERER F. & TROUVAIN J. (2015). Productions of /h/ in german : French vs. german speakers. In *Sixteenth Annual Conference of the International Speech Communication Association*.

ŻYGIS M. & PADGETT J. (2010). A perceptual study of polish fricatives, and its implications for historical sound change. *Journal of Phonetics*, **38**(2), 207–226.

Syllabe CVC et cycle mandibulaire : une étude articulatoire des asymétries. Le cas du vietnamien

Thi Thuy Hien Tran, Nathalie Vallée, Silvain Gerber
GIPSA-lab, Département Parole et Cognition, UMR 5216
CNRS – Université Grenoble Alpes, BP 25, 38040 Grenoble Cedex 9, France
thi-thuy-hien.tran@gipsa-lab.grenoble-inp.fr, nathalie.vallee@gipsa-lab.grenoble-inp.fr, silvain.gerber@gipsa-lab.grenoble-inp.fr

RESUME

Cette étude se situe dans le cadre d'un projet qui tente d'établir le lien entre asymétries phonétique et phonologique de la syllabe, plus spécifiquement le lien entre caractéristiques du geste mandibulaire et MOP, *Maximum Onset Principle*, principe phonologique empirique qui affecte les segments consonantiques à la position initiale de syllabe plutôt que finale. Plusieurs travaux antérieurs sur l'anglais américain ont montré l'existence d'asymétries au niveau des phases du cycle mandibulaire qui pourraient expliquer certaines tendances des structures syllabiques et notamment la structure canonique CV (plutôt que VC). Dans ce projet, une première étude sur le français a confirmé un patron d'asymétries mais inverse à celui trouvé pour l'anglais. Nous présentons les premiers résultats obtenus pour le vietnamien. Les résultats sont discutés dans deux cadres théoriques, *Frame/Content Theory* et *Articulatory Phonology*, le premier attribuant un rôle fondamental au geste mandibulaire dans la phonologie de la syllabe, le second ne lui concédant qu'un rôle secondaire.

ABSTRACT

CVC syllable and jaw cycle: an articulatory study of asymmetries. The case of Vietnamese.
The present paper is part of a project that tries to establish the link between phonological and phonetic asymmetries in syllable structures and more specifically the relationship between characteristics of jaw cycle and the MOP, *Maximum Onset Principle*, an empirical principle that affects consonantal segments to onset position rather than to coda position. Several previous works on American English have shown evidence of asymmetries within the jaw cycle, which could explain some syllabic structure tendencies, whereas a previous study on French showed a reverse asymmetrical pattern of jaw oscillation. We present here preliminary results on Vietnamese that we discuss in two theoretical frameworks namely *Frame/Content Theory* and *Articulatory Phonology*: the first one gives a fundamental role to the jaw in the shape of speech production, whereas the latter assigns a secondary role to the jaw in syllable phonology.

MOTS-CLES : Syllabe CVC, cycle mandibulaire, asymétries, vietnamien
KEYWORDS: CVC syllable, jaw cycle, asymmetries, Vietnamese

1 Introduction

The *Maximum Onset Principle* ou MOP est un principe empirique de syllabation qui place les consonnes en attaque de syllabe, position antéposée au noyau vocalique, plutôt qu'en position de coda, postposée au noyau vocalique (Kahn, 1976 ; Selkirk, 1982 ; Clements & Keyser, 1983 ;

Goldsmith, 1990 ; Blevins, 1995). Si bon nombre de travaux même expérimentaux en linguistique, psycholinguistique, traitement automatique des langues ont utilisé et utilisent le MOP, la question de sa nature est toujours en suspens.

La théorie *Frame/Content* (F/C) (MacNeilage, 1998, 2008) postule que la succession consonne-voyelle caractéristique de la syllabe canonique universelle CV est directement produite par le geste d'oscillation de la mandibule : une consonne est produite lorsque la mandibule est en position haute alors que la voyelle est réalisée lorsque la mandibule est en position basse. Le fondement de cette théorie selon laquelle une syllabe s'inscrit dans un cycle mandibulaire pourrait expliquer la nature du MOP qui favorise les structures syllabiques asymétriques. Cependant, la théorie F/C ne permet pas d'expliquer le caractère universel de la structure CV par rapport à la combinaison inverse VC, qui concerne moins de 5 % des syllabes des langues du monde (Rousset, 2004), et dont la structure s'insère tout aussi bien dans un cycle mandibulaire.

Une piste d'explication sur la nature du MOP pourrait résider dans les résultats des travaux de Kelso *et al*. (1985), Gracco (1994), Redford (1999) ou encore Redford & Donkelaar (2008) qui ont montré l'existence d'asymétries entre les deux phases du cycle de l'oscillation mandibulaire : la phase de remontée (fermeture) est observée plus rapide, plus courte, moins ample, plus raide (*stiffness*) que la phase d'abaissement (ouverture). Ces deux phases relèveraient directement des propriétés biomécaniques de la mandibule et auraient un impact sur l'articulation des segments dans les séquences sonores (successions de consonnes et voyelles). Le patron asymétrique du timing de l'oscillation mandibulaire trouvé dans ces travaux sur l'anglais américain pourrait rendre compte du MOP par le fait qu'il y aurait plus de délai temporel pour l'articulation des segments consonantiques lors de la phase d'ouverture que lors de la phase de fermeture. Ainsi, les phases asymétriques, si elles étaient vérifiées dans d'autres langues, pourraient expliquer plusieurs grandes tendances observées dans les langues du monde, les deux premières attribuées généralement au MOP : (1) la syllabe canonique universelle CV alors que la structure inverse VC est rare, (2) les clusters bien plus fréquents en attaque de syllabe, (3) la présence de consonnes avec articulation complexe (glottalisées, prénasalisées, labialisées, aspirées, clicks et autres occlusives doubles, etc.) en position pré-vocalique plutôt que post-vocalique, (4) une homorganicité du noyau vocalique trouvée plus fréquente avec la consonne en coda qu'avec la consonne en attaque (Vallée *et al*., 2009).

Cependant, une étude récente menée sur le français a mis en évidence au niveau du cycle mandibulaire, lors de la production de séquences CV.CVL.CVC (C= {/t/, /s/}, V=/a/, L=/l/), un patron d'asymétrie inverse à celui trouvé dans les études antérieures (voir plus haut) avec une phase d'ouverture plus courte, plus rapide et moins ample que la phase de fermeture (Vallée *et al*., 2014, 2015). Ce patron a été trouvé chez tous les locuteurs quel que soit le type de structure syllabique et quelle que soit la position de la syllabe dans la séquence, ainsi que dans la production de séquences VV (/aiaiaiaiai/).

Le français est une langue qui présente majoritairement des structures syllabiques ouvertes (73 % dont 54 % de CV et 8 % de V), les structures fermées représentant 26 % des syllabes (Vallée *et al*., 2001). Nous proposons ici d'étudier les caractéristiques du geste d'oscillation mandibulaire en vietnamien, langue austro-asiatique, isolante, tonale, monosyllabique sur le plan phonologique et de structures syllabiques majoritairement fermées (74.12 % dont 69.47 % de CVC) (Tran, 2011, p. 75). Le patron syllabique communément admis pour le vietnamien est $C_1(w)V(C_2)$ avec, entre parenthèses, les éléments facultatifs (Doan, 1999). Le noyau des syllabes à attaque vide est toujours précédé d'une occlusive glottale. C'est une langue qui connait une forte restriction de son inventaire consonantique en coda : à part les deux semi-consonnes /w j/, seules six consonnes sur vingt-trois sont admises dans cette position (/p t k m n ŋ/), dont /p/ jamais en attaque de syllabe. Au niveau syntactique, le vietnamien ne connait pas de processus de resyllabation relevant du fait qu'élisions,

enchaînements ou liaisons ne peuvent se produire en raison de la constitution et de la nature de la syllabe (Cao, 1985).

Dans la présente étude, les résultats obtenus sont discutés dans le contexte de la théorie Frame/Content selon laquelle l'origine de l'organisation syllabique de la parole est la caractéristique du cycle mandibulaire, mais aussi dans le cadre de la Phonologie Articulatoire (Browman & Goldstein, 1988, 1995, 2000) qui n'attribue à la mandibule qu'un rôle secondaire du fait qu'elle est porteur d'autres articulateurs (lèvre inférieure, langue) et n'aurait donc pas d'action directe pour la production de la parole. La Phonologie Articulatoire explique la prédominance de la structure CV par les caractéristiques phonétiques des consonnes et des voyelles impliquées dans la séquence dont les gestes respectifs sont produits en phase lorsque la consonne est pré-vocalique, alors qu'un geste consonantique post-vocalique est observé moins dépendant du geste vocalique et demande une coordination motrice plus complexe. La coordination naturellement en phase attaque-noyau donne plus de stabilité articulatoire à une structure CV (Browman & Goldstein, 1995) lui conférant sa caractéristique universelle et ce dès l'émergence de la parole (Goldstein, Byrd & Saltzman, 2006 ; Whalen, Giulivi, Goldstein, Nam & Levitt, 2011).

2 Procédure et méthode

2.1 Matériel

Les mouvements des articulateurs – mâchoire, langue, lèvres – ont été mesurés avec le système d'articulographie électromagnétique (EMA) AG2000 de la société Cartens grâce auquel il a été procédé à l'acquisition en 2D, à une fréquence de 200 Hz, de 5 bobines collées sur les articulateurs (mâchoire, lèvre inférieure, lèvre supérieure, apex et dos de la langue) et 2 bobines collées sur le plan de référence médio-sagittal du sujet afin de corriger les mouvements de la tête. Le signal acoustique de parole a été enregistré avec un enregistreur numérique stéréo PMD 670 de Marantz, micro directionnel C1000S d'AKG et numérisé à 44.1 KHz.

2.2 Corpus et participants

Le corpus vietnamien est constitué de logatomes de structures C_1V, /ʔ/VC_2, C_1VC_2 avec C_1 = /b d p s t z/, V = /a i/, C_2 = /p t/, ton modal montant B1-D1 *sắc*. Un autre corpus de 22 phrases contenant chacune des mots (simples ou composés) complète ces données, ainsi que la répétition d'un enchainement des 2 voyelles /aiaiaiaiai/. Des critères identiques pour C_1, V, C_2 et le ton ont été appliqués pour les mots simples et les premières syllabes de mots composés. Concernant la deuxième syllabe des mots composés $C_1VC_2.C_3VC_4$, nous avons neutralisé le ton (même ton *sắc*) et contrôlé la sonorité de C_3, attaque de la deuxième syllabe. La séquence /aiaiaiaiai/ a été enregistrée dans le but d'observer l'oscillation mandibulaire sans geste consonantique. Cinq locutrices natives du vietnamien de Hanoi (L_1 à L_5), âgées de 20 à 35 ans, ont participé à l'expérience. Les logatomes et phrases leur ont été présentés sous forme orthographe et enregistrés dans deux sessions à part. Les répétitions des logatomes comme des phrases ont été présentées dans un ordre aléatoire. Les résultats préliminaires présentés ci-après portent sur 5 répétitions des 12 logatomes de structure C/a/C, prononcés à un débit normal d'élocution : /bap/, /bat/, /dap/, /dat/, /pap/, /pat/, /sap/, /sat/, /tap/, /tat/, /zap/, /zat/. D'autres stimuli sont en cours d'analyse.

2.3 Mesures et analyses

Nous avons effectué, à l'heure actuelle, les mesures sur le mouvement de la mandibule. À partir de la segmentation semi-automatique avec EasyAlign et Praat, les séquences ont été extraites et

étiquetées avec un logiciel interne (TRAP) développé sous Matlab au GIPSA-lab par C. Savariaux. Sur les trajectoires des déplacements des différentes bobines ont été repérés automatiquement les minima et maxima à partir des passages par zéro de la courbe de vitesse de chacun des articulateurs. Les mesures effectuées sont les suivantes : (1) durée des phases d'abaissement (ouverture) et de remontée (fermeture) de la mandibule mesurées à partir des maxima d'ouverture et de fermeture qui correspondent aux points de passage par zéro de la courbe de vitesse ; (2) pic de vitesse et vitesse moyenne de chacune des phases ; (3) amplitude de chaque phase, qui correspond au déplacement vertical de la mandibule, estimée entre les maxima d'ouverture et de fermeture.

Nous souhaitons étudier les variations des variables réponses (durée, vitesse et amplitude de chacune des phases du cycle mandibulaire) et l'influence de deux facteurs sur celles-ci : phase (deux modalités) et logatome (douze modalités). Notre protocole ayant permis de recueillir plusieurs valeurs de variable réponse pour un même sujet, il ne nous garantit donc pas l'indépendance des observations. Notre choix s'est porté sur le modèle linéaire à effets mixtes et pour permettre de respecter l'hypothèse selon laquelle les résidus suivent une loi normale (condition d'application des modèles mixtes), nous avons choisi de transformer la variable réponse en son logarithme comme ce qui a été fait pour les données du français (Vallée *et al.*, 2014). Pour analyser la différence entre les deux modalités (phases d'ouverture et fermeture), à l'intérieur de chaque modalité type syllabique, nous appliquons la méthode de Hothorn, Bertz et Westfall (2008) qui permet de réaliser des comparaisons multiples de moyennes avec le modèle mixte en garantissant également que le risque de première espèce lié à la prise simultanée de toutes les décisions ne dépasse pas le seuil fixé à l'avance à 5 % en ajustant les p-values. La méthode a été appliquée aux données avec la fonction *glht* du package *multcomp* du logiciel R ainsi que la fonction *lsmeans* du package *lsmeans*.

3 Résultats

3.1 Durée

Globalement, les comparaisons multiples des durées moyennes entre phase d'ouverture et phase de fermeture pour chacun des cycles mandibulaires relevés dans les 12 logatomes montrent des différences significatives entre les deux phases chez tous les sujets, avec une fermeture plus longue que l'ouverture (Table 1, Figures 1 et 2).

	Estimate	Std. Error	z value	Pr(>|z|)
F - O | bap	0.3106	0.0419	7.4146	< 0.001
F - O | bat	0.2249	0.0419	5.3706	< 0.001
F - O | dap	0.3295	0.0428	7.7014	< 0.001
F - O | dat	0.3449	0.0419	8.2347	< 0.001
F - O | pap	0.3604	0.0419	8.6033	< 0.001
F - O | pat	0.2147	0.0419	5.1258	< 0.001
F - O | sap	0.3205	0.0419	7.6514	< 0.001
F - O | sat	0.259	0.0419	6.1838	< 0.001
F - O | tap	0.297	0.0428	6.9414	< 0.001
F - O | tat	0.2887	0.0428	6.7469	< 0.001
F - O | zap	0.3166	0.0428	7.3987	< 0.001
F - O | zat	0.3416	0.0428	7.9836	< 0.001

TABLE 1 – Estimations ponctuelles des différences de moyennes de durée (log) entre phase de fermeture (F) et phase d'ouverture (O) pour les cycles mandibulaires des 12 logatomes avec écart-type des différences, valeur de la statistique et p-value (hypothèse du test : F-O = 0).

La figure 2 comporte les valeurs des comparaisons multiples entre les différents types de cycles

mandibulaires (lignes horizontales), le point indiquant l'estimation ponctuelle de la différence et les parenthèses, les bornes de l'intervalle de confiance à 95 %. Tous les intervalles de confiance ne contiennent pas la valeur 0 indiquant une différence significative. En effet, la figure 1 montre que la durée de la phase de fermeture est largement supérieure à celle d'ouverture. Si on observe plus en détail, la remontée mandibulaire est plus longue pour une coda bilabiale (/bap/, /dap/, /pap/, /sap/, /tap/, /zap/) que pour une coda coronale (/bat/, /dat/, /pat/, /sat/, /tat/, /zat/). En ce qui concerne la phase d'ouverture, pour les 4 logatomes /bap/, /bat/, /pap/, /pat/, le relâchement d'une attaque bilabiale est plutôt stable quel que soit le type de consonne en coda. Alors que le relâchement d'une coronale suivie d'une labiale comme dans /dap/, /sap/, /zap/ est plus longue que le relâchement d'une coronale suivie d'une coronale comme dans /dat/, /sat/, /zat/. Aucune corrélation n'est observée entre durées d'ouverture et de fermeture pour toutes les locutrices (L_1 : r(58) = .25, p > .05 ; L_2 : r(57) = .16, p > .05 ; L_3 : r(57) = .06, p > .05 ; L_4 : r(46) = -.21, p > .05 ; L_5 : r(55) = -.19, p > .05).

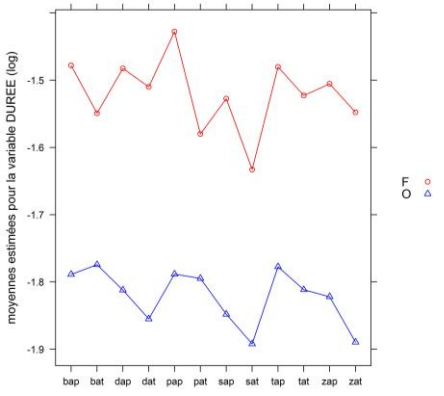


FIGURE 1 – Estimations des durées moyennes (log) des phases d'ouverture et de fermeture pour chaque cycle mandibulaire mesuré. Les fermetures sont significativement plus longues que les ouvertures.

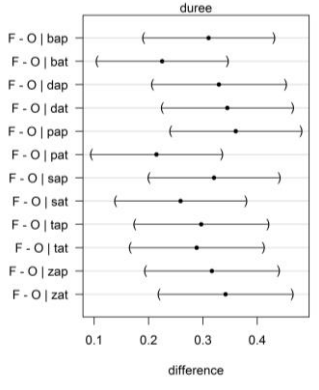


FIGURE 2 – Estimations ponctuelles des différences de moyennes entre durée de fermeture et durée d'ouverture (exprimées en log) et intervalles de confiance à 95 % pour chaque cycle mesuré.

3.2 Vitesse

Comme pour la durée, des différences significatives de la vitesse moyenne entre les deux phases ont été observées (Table 2, Figure 4) avec une ouverture plus rapide que la fermeture (Figure 3), sauf pour les cycles mandibulaires correspondant à la réalisation de /bat/ (p = .21) et de /pat/ (p = .99). La vitesse moyenne d'ouverture est généralement stable, sauf pour les attaques bilabiales sourdes (/pap/, /pat/) pour lesquelles le geste d'ouverture est mesuré moins rapide que dans le cas des attaques /b d s t z/ (Figure 3). Ceci est dû probablement au fait que /p/ n'est pas permis en vietnamien en attaque de syllabe. Le geste articulatoire pour /p/ semble donc plus contrôlé impliquant une vitesse moindre de la mandibule par rapport aux autres consonnes licites en attaque de syllabe. On observe plus de différences dans les vitesses de fermeture entre coronales et labiales : le geste de fermeture des coronales (/bat/, /dat/, /pat/, /sat/, /tat/, /zat/) est plus rapide que celui des labiales (/bap/, /dap/, /pap/, /sap/, /tap/, /zap/) (Figure 3). Pour la phase d'ouverture, phase du geste du relâchement de la consonne initiale vers la voyelle, un geste plus rapide des coronales que les labiales n'est pas observé.

Une corrélation entre vitesses moyennes des deux phases est également remarquée pour chaque sujet (L_1 : r(58) = .28, p < .05 ; L_2 : r(57) = .39, p < .05 ; L_4 : r(46) = .83, p < .05 ; L_5 : r(55) = .35, p < .05), sauf pour la locutrice L_3 (r(57) = .01, p > .05) qui présente un débit de parole plus rapide que les autres locutrices (Figure 6).

	Estimate	Std. Error	z value	Pr(>|z|)
F - O | bap	-0.3202	0.0522	-6.1382	< 0.001
F - O | bat	-0.1211	0.0522	-2.3218	0.2176
F - O | dap	-0.3593	0.0533	-6.7443	< 0.001
F - O | dat	-0.2529	0.0522	-4.8496	< 0.001
F - O | pap	-0.241	0.0522	-4.6212	< 0.001
F - O | pat	0.0404	0.0522	0.7748	0.999
F - O | sap	-0.6105	0.0522	-11.7051	< 0.001
F - O | sat	-0.3483	0.0522	-6.6771	< 0.001
F - O | tap	-0.4438	0.0533	-8.3297	< 0.001
F - O | tat	-0.2726	0.0533	-5.1157	< 0.001
F - O | zap	-0.5489	0.0533	-10.3015	< 0.001
F - O | zat	-0.4329	0.0533	-8.1253	< 0.001

TABLE 2 – Estimations ponctuelles des différences de moyennes de vitesses (log) entre phase de fermeture (F) et phase d'ouverture (O) pour les cycles mandibulaires correspondant aux 12 logatomes avec écart-type des différences, valeur de la statistique et p-value (hypothèse du test : F-O=0).

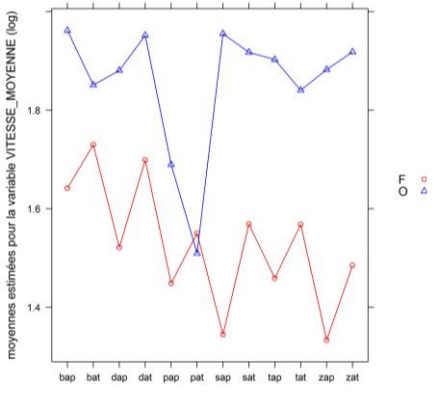


FIGURE 3 – Estimations des vitesses moyennes (log) des phases d'ouverture et de fermeture pour chaque cycle mandibulaire. Sauf pour /pat/ et /bat/, les ouvertures sont significativement plus rapides.

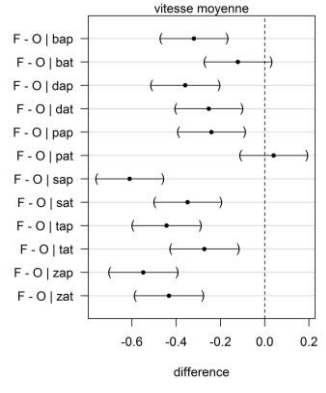


FIGURE 4 – Estimations ponctuelles des différences de moyennes entre vitesse de fermeture et vitesse d'ouverture (log) et intervalles de confiance à 95 % pour chaque cycle mandibulaire.

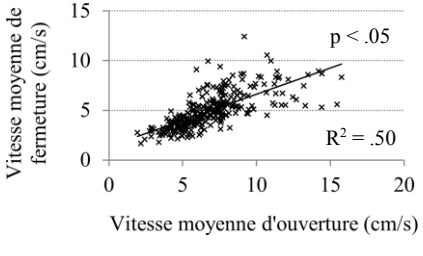


FIGURE 5 – Corrélation entre vitesses moyennes des phases d'ouverture et de fermeture du cycle mandibulaire pour tous les sujets.

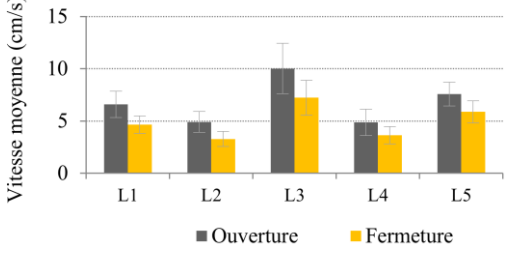


FIGURE 6 – Vitesse moyenne des phases d'ouverture et de fermeture du cycle mandibulaire chez chaque locutrice (cm/s).

La même observation est valable pour les pics de vitesse. Ces pics, mesurés et moyennés pour chaque logatome et chaque locutrice, suivent les mêmes patrons déjà présentés pour les vitesses moyennes, avec un pic d'ouverture significativement plus rapide que celui de fermeture ($p < .001$), sauf pour les cycles mandibulaires correspondant à la réalisation de /bat/ ($p = .99$) et /pat/ ($p = .70$).

3.3 Amplitude

L'amplitude (déplacement vertical de la mandibule), ne montre pas de patron régulier entre les phases d'ouverture et les phases de fermeture du cycle mandibulaire. On relève une amplitude significativement plus grande à l'ouverture qu'à la fermeture (Figure 7) pour les cycles avec une consonne fricative coronale en attaque et une consonne labiale en coda /sap/ et /zap/ ($p < .001$) et le contraire, une amplitude significativement plus petite à l'ouverture qu'à la fermeture pour le logatome /pat/ ($p < .001$). On note chez chaque locutrice une corrélation forte entre amplitude d'ouverture et amplitude de fermeture (L_1 : $r(58) = .33$, $p < .05$; L_2 : $r(57) = .47$, $p < .05$; L_3 : $r(57) = .61$, $p < .05$; L_4 : $r(46) = .74$, $p < .05$; L_5 : $r(55) = .87$, $p < .05$).

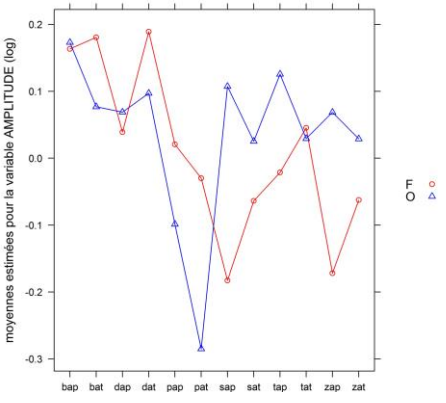


FIGURE 7 – Estimations du déplacement vertical de la mandibule (log) dans les phases d'ouverture et de fermeture pour chaque syllabe.

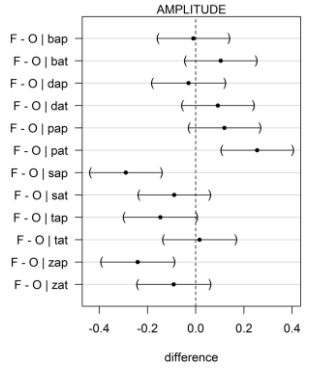


FIGURE 8 – Estimations ponctuelles des différences de moyenne entre amplitudes de fermeture et d'ouverture (log) et intervalles de confiance à 95 % pour chaque syllabe. Seules les différences dans /sap/, /zap/, /pat/ sont significatives.

3.4 Comparaison avec le français

Ces résultats préliminaires sur des séquences CVC du vietnamien nous ont permis de comparer nos résultats avec les données déjà obtenues auprès de cinq locutrices du français pour le même type de séquence (Vallée *et al*., 2014, 2015). Des tendances similaires sont observées pour les deux langues, avec des phases d'ouverture significativement plus courtes et plus rapides que les phases de fermeture. Les productions de la séquence vocalique /aiaiaiaiai/ par les locutrices vietnamiennes et françaises présentent les mêmes profils d'asymétries que les séquences avec geste consonantique. Cette séquence a été intégrée au corpus dans le but d'observer l'oscillation mandibulaire dans un contexte considéré comme plutôt stable en parole répétée (Kelso *et al*., 1985). Ce résultat consolide donc nos observations faites sur les séquences avec geste consonantique.

4 Discussion et conclusion

Notre étude du geste mandibulaire dans des productions de syllabe CVC en vietnamien montre l'existence chez tous les locuteurs de patrons asymétriques du cycle d'oscillation, avec une phase d'ouverture (abaissement de la mandibule) plus courte, plus rapide que la phase de fermeture, ce qui suggère qu'ouverture et fermeture ne sont pas soumises aux mêmes contraintes. Les corrélations

positives trouvées systématiquement entre les deux phases pour les facteurs vitesse moyenne, pic de vitesse et amplitude, indiquent que les phases d'ouverture et de fermeture ne sont pas des actions indépendantes, confirmant Gracco (1994). Les mêmes observations avaient été faites dans l'étude antérieure de Vallée *et al*. (2014) à propos de transitions VV produites par des locuteurs vietnamiens. Le mode articulatoire des segments (plosive *vs*. fricative) et leur distribution (position initiale *vs*. finale de syllabe) n'impactent pas les patrons d'asymétries observés dans notre étude pour les facteurs durée et vitesse. Nos résultats montrent aussi que les valeurs de vitesse et d'amplitude du geste mandibulaire sont influencées par des caractéristiques intrinsèques (lieu articulatoire labial *vs*. coronal) et extrinsèques des segments (distribution licite *vs*. illicite dans la langue observée dans les syllabes /p/VC).

Ces résultats sont similaires à ceux obtenus pour le français par Vallée *et al*. (2014) et excluent pour l'instant toute explication liée à la structure syllabique de base des deux langues. Le français est une langue à structure de base CV (Rousset, 2004) ; le vietnamien est une langue majoritairement CVC pour laquelle l'étude de Tran (2011) a montré que les durées acoustiques des plosives étaient similaires en attaque et en coda. Ce résultat de Tran (2011) ne reflète pas les patrons d'asymétries du geste mandibulaire observé dans notre étude, suggérant une indépendance entre durée acoustique des plosives et durée des phases mandibulaires et donc l'absence de contrainte temporelle du geste mandibulaire sur l'articulation des segments. Notre résultat contredit la corrélation trouvée par Redford (1999) entre durée des segments et durée des phases, avec une durée plus courte des segments en coda observée lors de la phase de fermeture car plus courte que la phase d'ouverture. Plus généralement, à ce stade, nos résultats sont en contradiction avec ceux des études antérieures sur l'anglais-américain (Kelso *et al.*, 1985 ; Gracco, 1994) et l'extension de la théorie *Frame/ Content* proposée par Redford (1999) qui montraient des patrons d'asymétries inverses supposés être la conséquence des propriétés biomécaniques de la mandibule.

Cependant, une phase d'ouverture plus courte et plus rapide est conciliable avec une relation de phasage inter-geste attaque-noyau plus stable (Byrd, 1996) et un degré de cohésion plus important des gestes (Browman & Goldstein, 2000) par rapport à noyau-coda. Les patrons d'asymétries du cycle mandibulaire relevés dans notre étude sont cohérents avec les propositions de la Phonologie Articulatoire : (1) « *in phase coupling* » pour CV produite lors de la phase d'ouverture ; (2) « *time-independent even in an anti-phase relation* » pour VC réalisée lors de la phase de fermeture et cohérente avec une dépendance de timing et un chevauchement des gestes noyau-coda moins importants qu'attaque-noyau (Browman & Goldstein, 1988).

Dans le prolongement de cette étude, est en cours une analyse du phasage du geste mandibulaire avec le geste labial ou lingual qui montre, par exemple pour la syllabe /pat/, que l'écart temporel entre l'atteinte de la cible /p/ (point maximal atteint par la lèvre inférieure pour réaliser l'occlusion) et la cible /a/ (maximum d'ouverture de la mandibule) est plus long que l'écart temporel entre la cible /a/ et la cible /t/ (maximum atteint par l'apex pour l'occlusion de /t/). On retrouve ici l'asymétrie mise en évidence dans les études antérieures sur l'anglais américain mais mesurée entre le cycle mandibulaire et le cycle de la langue ou des lèvres. Ce constat ne peut que nous inciter à poursuivre nos investigations multi-locuteurs et inter-langues. Dans une perspective proche, des comparaisons avec les corpus de phrases en français et en vietnamien sont envisagées. L'acquisition de données nouvelles sur l'anglais américain est également prévue.

Remerciements

Cette étude fait partie du projet ANR-10-BLAN-1916 *APPSy*. Nous remercions vivement Christophe Savariaux et Quentin Tura pour leur assistance précieuse dans ce projet.

Références

BLEVINS J. (1995). The syllable in Phonological Theory. In Goldsmith J. A. (eds.): *Handbook of Phonological Theory*, 206-235. Blackwell Publishers: Oxford.

BROWMAN C. P. & GOLDSTEIN L. M. (1988). Some notes on syllable structure in articulatory phonology. *Phonetica* 45, 140-155.

BROWMAN C. P. & GOLDSTEIN L. M. (1995). Gestural syllable position effects in American English. In BELL-BERTI F. & RAPHAEL L. J.: *Producing Speech: Contemporary Issues. For Katherine Safford Harris*, 19-33. AIP Press: New York.

BROWMAN C. P. & GOLDSTEIN L. M. (2000). Competing constraints on intergestural coordination and self-organization of phonological structures. *Bulletin de la Communication Parlée* 5, 25-34.

BYRD D. (1996). A phase window framework for articulatory timing. *Phonology*, 13(02), 139-169.

CAO X. H. (1985). *Phonologie et linéarité. Réflexions critiques sur les postulats de la phonologie contemporaine*. Paris: SELAF.

CLEMENTS N. & KEYSER S. J. (1983). *CV Phonology: A generative theory of the syllable*. MIT Press.

ĐOÀN T. T. (1999). *Ngữ âm tiếng Việt* (Tr. La phonétique du vietnamien). Hanoi: Nhà xuất bản Đại học Quốc gia Hà Nội (Maison d'édition de l'Université Nationale de Hanoi).

GOLDSTEIN L. M., BYRD D. & SALTZMAN E. L. (2006). The role of vocal tract gestural action units in understanding the evolution of phonology. In ARBIB M. (eds.): *From action to language: The mirror neuron system*, 215–249. Cambridge University Press: Cambridge.

GRACCO V. L. (1994). Some organizational characteristics of speech movement control. *Journal of Speech and Hearing Research* 37, 4-27.

HOTHORN T., BERTZ F. & WESTFALL P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal* 50(3), 346-363.

KAHN D. (1976). *Syllable-based Generalizations in English Phonology*. Indiana University Linguistics Club: Bloomington.

KELSO J. A. S., VATIOKIOTIS-BATESON E., SALTZMAN E. L. & KAY B. (1985). A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics and dynamic modeling. *JASA* 77, 266-280.

MACNEILAGE P. F. (1998). The Frame/Content Theory of Evolution of Speech Production. *Behavioral and Brain Sciences* 21, 499-511.

MACNEILAGE P. F. (2008). *The origin of speech*. Oxford University Press: Oxford, England.

REDFORD M. A. (1999). *An Articulatory Basis for the Syllable*. PhD thesis, The University of Texas: Austin.

REDFORD M. A. & VAN DONKELAAR P. (2008). Jaw cycles and linguistic syllables in adult English. In DAVIS B.L. & ZAJDO K. (eds.): *The Syllable in Speech Production: Perspectives on the Frame/Content Theory*, 355-76. Taylor & Francis: London.

ROUSSET I. (2004). *Structures syllabiques et lexicales des langues du monde : données, typologies, tendances universelles et contraintes substantielles*. Thèse de doctorat, Université Stendhal Grenoble.

SELKIRK E. (1982). The syllable. In VAN DER HULST H. & SMITH N. (eds.): *The structure of phonological representations*, 337-383. Foris Publications: Dordrecht.

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, Université de Grenoble.

VALLEE N., ROUSSET I., BOË L. J. (2001) Des lexiques aux syllabes des langues du monde. Typologies, tendances et organisations structurelles. *Linx* 45 : 37-50.

VALLEE N., ROSSATO, S., ROUSSET I. (2009). Favoured syllabic patterns in the world's languages and sensorimotor constraints. In PELLEGRINO F., MARSICO E., CHITORAN I. & COUPÉ C. (eds.): *Approaches to Phonological Complexity*, 111-139. Mouton de Gruyter: Berlin.

VALLEE N., TRAN T. T. H, ROSSATO S., MAIRANO P. (2014). Structures syllabiques et caractéristiques du cycle mandibulaire : une étude articulatoire des asymétries. *XXXe Journées d'Etudes sur la Parole* (JEP 2014), Le Mans

VALLEE N., TRAN T. T. H, ROSSATO S., MAIRANO P., GERBER S. (2015). Why do syllable onsets attract consonant(s)? *Italian Journal of Linguistics*, pp. 133-160, volume 27, issue 1, Scuola Normale Superiore.

WHALEN D. H., GIULIVI S., GOLDSTEIN L. M., NAM H. & LEVITT A.G. (2011). Response to MacNeilage and Davis and to Oller. *Language Learning and Development* 7(3), 243–249.

De l'utilisation de descripteurs issus de la linguistique computationnelle dans le cadre de la synthèse par HMM

Sébastien Le Maguer[1] Bernd Möbius[1] Ingmar Steiner[1,2] Damien Lolive[3]
(1) Computational Linguistics and Phonetics, Saarland University, Saarbrücken, Allemagne
(2) DFKI, Saarbrücken, Allemagne
(3) IRISA, Lannion, France
{slemaguer|moebius|steiner}@coli.uni-saarland.de
damien.lolive@irisa.fr

RÉSUMÉ

Durant les dernières décennies, la modélisation acoustique effectuée par les systèmes de synthèse de parole paramétrique a fait l'objet d'une attention particulière. Toutefois, dans la plupart des systèmes connus, l'ensemble des descripteurs linguistiques utilisés pour représenter le texte reste identique. Plus specifiquement, la modélisation de la prosodie reste guidée par des descripteurs de bas niveau comme l'information d'accentuation de la syllabe ou bien l'étiquette grammaticale du mot. Dans cet article, nous proposons d'intégrer des informations basées sur la prédictibilité d'un évènement (la syllabe ou le mot). Plusieurs études indiquent une corrélation forte entre cette mesure, fortement présente dans la linguistique computationnelle, et certaines spécificités lors de la production humaine de la parole. Notre hypothèse est donc que l'ajout de ces descripteurs améliore la modélisation de la prosodie. Cet article se focalise sur une analyse objective de l'apport de ces descripteurs sur la synthèse HMM pour la langue anglaise et française.

ABSTRACT

Toward the use of information density based descriptive features in HMM based speech synthesis

Over the last decades, acoustic modeling for speech synthesis has been improved significantly. However, in most systems, the descriptive feature set used to represent annotated text has been the same for many years. Specifically, the prosody models in most systems are based on low level information such as syllable stress or word part-of-speech tags. In this paper, we propose to enrich the descriptive feature set by adding a linguistic measure computed from the predictability of an event, such as the occurrence of a syllable or word. By adding such descriptive features, we assume that we will improve prosody modeling. This new feature set is then used to train prosody models for speech synthesis. This paper focuses on an objective analysis of the influence of these descriptive features on the synthesis achieved in English and French.

MOTS-CLÉS : Synthèse de la parole paramétrique, densité d'information, descripteurs linguistiques.

KEYWORDS: Parametric speech synthesis, information density, descriptive features.

1 Introduction

Durant ces dernières années, la popularité de la synthèse paramétrique a pris de l'ampleur au point de devenir l'une des méthodologies standards de la synthèse text-to-speech (TTS). De la synthèse par hidden Markov model (HMM) (Zen & Toda, 2005) aux réseaux de neurones profonds (Zen *et al.*, 2013), l'effort de recherche s'est principalement focalisé sur la modélisation acoustique.

Toutefois, l'ensemble de ces systèmes se basent sur le même ensemble de descripteurs linguistiques et prosodiques pour prédire les paramètres acoustiques. Ainsi, la majorité des systèmes basés sur HMM utilisent un jeu de descripteurs dérivés du jeu proposé dans Tokuda *et al.* (2002). Peu de travaux se focalisent sur l'intégration de nouveaux descripteurs linguistiques et prosodiques. Parmi celles ci, nous pouvons citer l'utilisation de « word embeddings » Wang *et al.* (2015) ou bien l'utilisation d'informations syntaxiques enrichies (Obin *et al.*, 2010).

Dans Le Maguer *et al.* (2016), nous avons proposé d'intégrer de nouveaux descripteurs, issus du domaine de la « densité d'information » (Information Density). Les résultats montrent que l'utilisation de tels descripteurs améliore la synthèse effectuée. Ces descripteurs sont basés sur l'évaluation de l'imprédictibilité d'un évènement ; notion répandue en linguistique computationnelle. Ils sont obtenus en utilisant un modèle de langue ce qui procure plusieurs avantages. Tout d'abord, ils sont simples à obtenir à partir d'un texte et peuvent être utilisés pour aboutir à des descripteurs de plus haute abstraction linguistique. Ils peuvent également être utilisés pour la synthèse classique ou bien la synthèse incrémentale (Baumann & Schlangen, 2012; Pouget *et al.*, 2015).

Dans le présent article, nous proposons d'appliquer cette méthodologie pour une synthèse HMM pour le français. Nous proposons également d'analyser l'impact de ces descripteurs sur les modèles appris et de comparer cet impact aux résultats obtenus pour l'anglais. Cette comparaison est rendue possible par le fait que le jeu de descripteurs que nous avons utilisé pour les deux langues est similaire.

Ainsi, cet article est organisé de la manière suivante : la section 2 présente et justifie l'utilisation de tels descripteurs. La section 3 décrit le protocole d'évaluation mis en place pour analyser l'influence de ces descripteurs sur la synthèse. La dernière section (4) présente les résultats de l'évaluation objective pour le français et l'anglais.

2 Descripteurs linguistiques basés sur la densité d'information

Reposant sur la théorie de l'information proposée par Shannon (1948), Hale (2001) introduit le concept d'imprédictibilité (*surprisal*), associé à la densité d'information, au sein du domaine de la linguistique computationnelle. Il repose sur une modélisation n-gramme et est défini par l'équation suivante :

$$Surp(U_i) = -log_2(P(U_i|U_{i-1}..U_{i-1-N})) \quad (1)$$

où U_i correspond à l'unité analysée et $U_{i-1}..U_{i-1-N}$ sont les N unités précédentes. N est le paramètre du modèle et doit être défini.

L'utilisation d'un tel concept repose sur le résultat d'études issues du domaine de la psycho-linguistique. En effet, la prédictibilité d'un mot est fortement corrélée avec l'effort nécessaire pour prononcer ce mot (Smith & Levy, 2013). Une corrélation analogue a été mise en avant avec la prédictibilité d'une syllabe (Jaeger, 2010). Ainsi, notre hypothèse est que l'introduction de descrip-

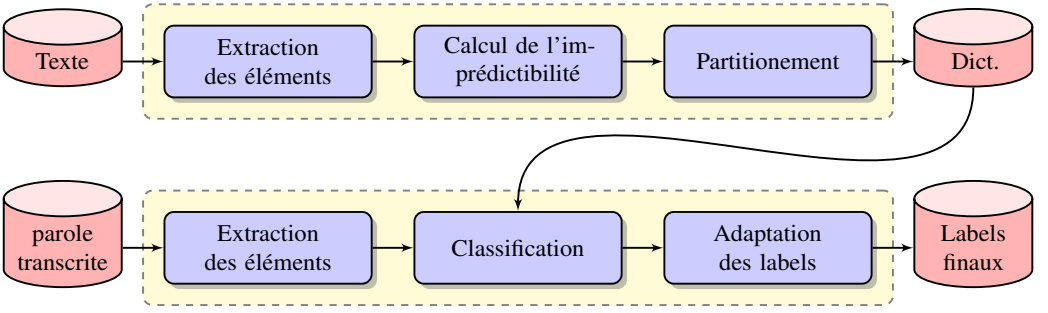


FIGURE 1 – Génération des descripteurs basés sur la densité d'information. À partir du corpus de texte, nous souhaitons extraire un dictionnaire qui associe un cluster d'imprédictibilité à chaque élément (syllabe ou mot). Tout d'abord, nous extrayons les éléments en utilisant des outils TALN et nous calculons leur imprédictibilité via (1). Ensuite, nous partitionnons l'espace obtenu afin de déterminer une échelle allant du plus prédictible (valeur à 0) au moins prédictible. L'étape finale consiste à associer à chaque élément une partition en se basant sur l'imprédictibilité de cet élément. À partir du corpus de parole, nous extrayons le même type d'élément et, grâce au dictionnaire, déterminons la partition associée. Enfin, nous adaptons les fichiers de labels afin de les rendre compatibles avec le système de synthèse.

teurs linguistiques basés sur la prédictibilité devrait améliorer la synthèse et plus spécifiquement la modélisation de la prosodie. En effet, concernant la synthèse par HMM, la modélisation du spectre est davantage contrôlée par les informations phonologiques (Le Maguer *et al.*, 2013).

Toutefois, l'utilisation de l'imprédictibilité a deux problèmes principaux. Le premier est que, pour obtenir des statistiques valides, il nous faut disposer d'un corpus textuel beaucoup plus large que les corpus de paroles généralement utilisés. Le second est que la prédictibilité est un descripteur à valeurs continues alors que les systèmes TTS standards reposent sur l'utilisation de descripteurs à valeurs discrètes.

2.1 Processus général

Afin de palier ces problèmes, nous proposons l'utilisation du processus décrit dans la figure 1.

Tout d'abord, nous avons décidé d'utiliser deux corpora : un corpus de parole et un corpus de textes beaucoup plus large que le précédent. À partir du corpus de texte, nous déterminons l'imprédictibilité des évènements analysés et nous générons un dictionnaire de classes d'imprédictibilité. Ces classes sont définies en utilisant un algorithme de partitionnement, dans notre cas celui des k-moyennes. Ceci nous permet d'obtenir une approximation discrète des valeurs continues.

2.2 Descripteurs associés à la syllabe

Puisque nous utilisons deux corpora, pouvant être obtenus par différents outils TALN, nous proposons d'utiliser une représentation reposant sur l'alphabet phonétique international (International Phonetic Alphabet (IPA)) pour les phonèmes constituant chaque syllabe. Il est possible d'ajouter des infor-

mations plus spécifiques. Toutefois, dans cette étude, nous utilisons uniquement une représentation IPA.

2.3 Descripteurs associés au mot

Afin de représenter les mots, nous devons rester le plus possible proche du texte. Néanmoins, pour obtenir une représentation adaptée, il est nécessaire de nettoyer ce texte. Nous avons procédé de la manière suivante :

— tous les signes de ponctuation sont supprimés ;
— un indicateur typographique (le symbole #) est inséré à la fin de chaque paragraphe ;
— tous les mots sont convertis en minuscule.

Les indicateurs sont traités comme des mots à part entière. Ils ont été insérés à la fin de chaque paragraphe car nous faisons l'hypothèse qu'un paragraphe est conceptuellement homogène (le paragraphe ne traite que d'un seul « sujet »). L'apprentissage de modèles de parole est généralement basé sur des énoncés globalement courts. Ainsi, l'utilisation d'indicateurs est plus cohérente que l'utilisation d'une valeur non-définie en début de chaque énoncé. En effet, ajouter un indicateur permet aux arbres de décision de prendre en compte des partitions qui auraient été autrement ignorées.

3 Protocole expérimental

3.1 Corpus anglais

Le corpus anglais est issu du challenge Blizzard 2013 (King & Karaiskos, 2013). Il est composé de 83 livres audio lus par un locuteur féminin. De ce jeu de données, 2 corpora sont extraits : le *corpus de textes* et le *corpus de parole*. Le *corpus de textes* est composé de l'ensemble du corpus anglais excepté le livre « Black Beauty ». Cela correspond à 82 livres, 2 298 055 syllabes et 1 973 368 mots. Le *corpus de parole* est composé d'environ une heure de parole (∼470 énoncés) extraits de « Black Beauty » soit 13 522 syllabes et 7038 mots. Pour chaque corpus, les frontières de syllabes ont été obtenues grâce au système MaryTTS (Schröder & Trouvain, 2003) (version 5.2).

3.2 Corpus français

Le corpus français est constitué du roman « À la recherche du temps perdu » de Marcel Proust. De manière analogue au corpus anglais, 2 corpora sont extraits de ce jeu de données. Le *corpus de textes* est composé de l'ensemble du corpus français excepté le tome « Albertine disparue ». Cela correspond à 6 tomes, 1 806 672 syllabes et 1 005 492 mots. Le *corpus de parole* est composé d'environ une heure de parole (∼520 énoncés) extraits de « Albertine disparue » soit 15 088 syllabes et 10 051 mots. Pour chaque corpus, les frontières de syllabes ont été obtenues par un système par règles.

3.3 Analyse de l'imprédictibilité sur les corpus

Pour déterminer l'imprédictibilité, nous utilisons des trigrammes de mots et des trigrammes de syllabes.

Pour le corpus de parole anglais, l'ensemble des trigrammes des mots est présent dans le corpus de textes. En revanche, pour le corpus français, 33 % sont absents. Pour les trigrammes de mots, environ 40 % des instances du corpus de parole sont absentes du corpus de textes pour l'anglais ; contre 80 % pour le corpus français. En considérant les unités qui sont présentes dans les deux corpus, pour chaque langue, la figure 2 illustre la distribution de la prédictibilité des syllabes et mots distincts.

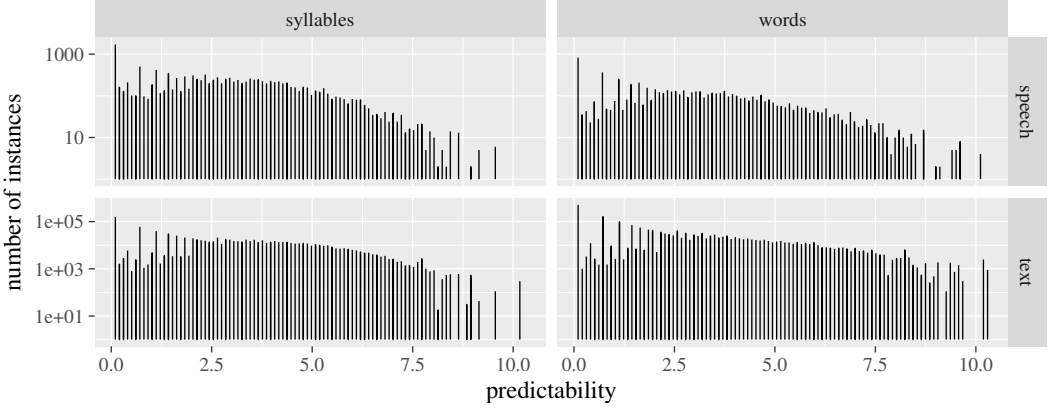


FIGURE 2 – Distribution de l'imprédictibilité des syllabes et des mots pour les corpus de textes et de parole

L'ensemble des distributions respectent le même schéma. Une imprédictibilité de 0 indique que l'évènement est qualifié intégralement par son contexte. En effet, les séquences $U_{i-1}...U_{i-N}$ et $U_i...U_{i-N}$ n'apparaissant qu'une seule fois dans le corpus ; la détection de $U_{i-1}...U_{i-N}$ implique forcément l'apparition de U_i.

Ensuite, nous avons les évènements fréquents avec une valeur d'imprédictibilité faible. Ces évènements correspondent aux schémas linguistiques courants tels que « one of those » pour l'anglais. L'utilisation de trigrammes, comparativement à une modélisation n-gramme plus large, a accru le nombre de ces évènements. Enfin, le nombre d'occurrences des éléments décroits au fur et à mesure que leur imprédictibilité augmente.

Le dernier paramètre du modèle correspond au nombre de partitions utilisées lors de la phase de partitionnement. Pour les expériences effectuées, nous avons utilisé 9 partitions ce qui correspond à la configuration par défaut de l'outil utilisé. En conséquence, l'imprédictibilité d'un évènement est définie sur une échelle 0 à 8, avec 0 indiquant une unité intégralement définie par son contexte, et 8 une unité dont les occurrences sont les plus inattendues.

3.4 Analyse des descripteurs linguistiques

Afin d'analyser l'influence des descripteurs proposés, nous avons défini trois jeux de descripteurs *baseline*, *pred_syllable*, and *pred_all*.

Le jeu *baseline* correspond au jeu proposé par Tokuda *et al.* (2002) pour l'anglais et par Le Maguer *et al.* (2013) pour le français. Les jeux sont similaires pour les deux langues mais obtenus via des outils logiciels différents. Dans les deux cas, l'information de stress associée à la syllabe a été écartée et il n'y a pas d'équivalent d'étiquette ToBI pour le jeu français.

Les deux autres configurations intègrent, respectivement, l'information d'imprédictibilité de la syllable puis cette même information pour la syllabe et le mot.

3.5 Système de synthèse

Afin d'analyser les descripteurs proposés, le système de synthèse utilisé est la configuration standard du système HMM based speech synthesis system (HTS) (Zen & Toda, 2005) (version 2.3) avec le vocodeur STRAIGHT (Kawahara *et al.*, 1999).

4 Analyse des modèles obtenus

Lors de notre première analyse pour l'anglais, présenté dans Le Maguer *et al.* (2016), nous avons constaté que l'évaluation par distance n'indiquait pas d'amélioration de la similarité. En revanche les évaluations subjectives montrent une amélioration perceptible de la synthèse.

Ainsi, afin d'analyser l'influence de ces descripteurs sur la synthèse, nous allons nous focaliser sur l'évolution de la structure des arbres de décision suivant les différentes configurations.

Pour les arbres de décisions du système HTS, les nœuds correspondent à des propriétés associées aux descripteurs linguistiques et les feuilles correspondent aux modèles eux mêmes. En conséquence, en considérant le corpus d'apprentissage et un arbre de décision, il est possible de déterminer l'importance accordée par le système à un descripteur. Enfin, nous avons groupé les descripteurs par catégorie afin d'aboutir à un analyse plus globale.

Dans notre cas, nous considérons les catégories suivantes :

— p$\{1, 3, 5\}$ pour la taille de la fenêtre à l'horizon du phonème (monophone, triphone, quinphone) ;
— $\{syl, word, phrase\}$-position pour les informations de position (e.g. position de la syllabe courante dans le mot courant) correspondant au niveau linguistique donné ;
— $\{syl, word, phrase\}$-prosody pour les descripteurs associés à la prosodie (accentuation de la syllabe, part-of-speech (POS) associé au mot) ;
— énoncé pour l'ensemble des informations de comptage global (nombre total de syllabes, de mots et de groupes prosodiques).

Nous y ajoutons deux catégories, $\{syl, word\}$-predictability, qui correspondent aux descripteurs proposés.

	Anglais			Français		
catégories	**baseline**	**pred_syl.**	**pred_all**	**baseline**	**pred_syl.**	**pred_all**
p1	1648	1615	1594	3202	3135	3152
p3	2174	2037	1175	2495	1170	1247
p5	0	0	694	0	892	897
syl-position	**5799**	**5652**	**4056**	**13 034**	**9405**	**9820**
syl-prosody	98	128	183	54	21	24
syl-predictability	0	1657	**4163**	0	**10 773**	**9884**
word-position	**7836**	**6787**	**4202**	**13 747**	**9235**	**10 406**
word-prosody	2928	2817	2188	**18 496**	**13 605**	**14 438**
word-predictability	0	0	**7834**	0	0	4548
phrase-position	1184	1573	802	2833	2878	3147
phrase-prosody	**8723**	**8323**	**5892**	0	0	0
utterance	**7260**	**7429**	**6799**	**16 421**	**13 922**	**15 139**

TABLE 1 – Analyse de l'évolution de l'arbre de décision du F0 associé à l'état central du HMM en utilisant les labels du corpus de parole. Chaque label est passé à travers l'arbre. À chaque nœud de l'arbre est associée une catégorie et chaque fois qu'un nœud est atteint le compteur de la catégorie correspondante est mis à jour. Les catégories les plus utilisées ont été surlignées.

Enfin, nous avons appliqué la normalisation suivante :

$$\text{Freq\_occupation (noeud)} = \frac{\text{Freq\_occupation (noeud)}}{\text{Nb\_classes(descripteur(noeud))}} \tag{2}$$

Cela permet d'éviter d'accorder une importance injustifiée à une catégorie simplement parce que le nombre de propriétés associées à celle-ci est plus importante que pour les autres catégories.

En considérant les différents arbres, l'information de prédictibilité a un impact limité sur la modélisation du spectre et de l'apériodicité. En effet, pour ces arbres, les informations principalement utilisées correspondent aux étiquettes des phonèmes. En revanche, Les arbres de décision associés à la durée et au F0 sont impactés par cette catégorie de descripteurs. De plus, les arbres associés à ces deux paramètres ont une structure similaire.

En conséquence, nous analysons uniquement les résultats obtenus pour l'arbre de décision de l'état central du HMM pour le F0. Ces résultats sont présentés dans le tableau 1.

Tout d'abord, en comparant les résultats pour le français et pour l'anglais, on peut remarquer une différence significative du nombre d'utilisation de nœuds (maximum 8723 pour l'anglais, 18496 pour le français). Ceci s'explique par la différence de durée entre les corpus mais également par un débit différent des locuteurs. En effet, le corpus anglais contient environ 20 % de segments en moins.

En se focalisant sur la configuration par défaut (*baseline*), la principale différence entre les deux langues repose sur l'utilisation du tag grammatical et les informations de position à l'horizon du groupe de souffle. Ceci s'explique par le fait que les questions (et valeurs possibles des descripteurs) associées au tag grammatical sont plus précises et nombreuses en français. En effet, pour l'anglais,

il n'y aucune distinction entre les tags « signifiants ». En revanche, le jeu de descripteurs français distingue les verbes, noms,...

Étonnamment, dans les deux cas, l'information d'accentuation n'est utilisée que pour affiner le modèle ; elle n'est pas considérée par HTS, pour ces jeux de données, comme une propriété fondamentale du F0. Ceci peut s'expliquer par le fait que l'information d'accentuation peut être capturée par l'information de position associée à la syllabe.

Prendre en compte l'imprédictibilité à l'horizon de la syllabe n'impacte pas fondamentalement les modèles pour le corpus anglais. En revanche, pour le corpus français, ce descripteur devient l'un des plus utilisé. En comparant la répartition d'occupation de nœuds pour le français, nous constatons que les informations de position à l'horizon de la syllabe et du mot sont fortement impactés par l'introduction de l'imprédictibilité à l'horizon de la syllabe. Cet impact est également visible pour l'anglais mais de manière plus limité. L'imprédictibilité étant déterminée sur un contexte de trois syllabes, il n'est pas étonnant d'aboutir à un tel résultat. En effet, l'information du nombre de syllabes contenues dans un mot est implicitement encodé dans l'information d'imprédictibilité de part l'introduction de ce contexte.

Enfin, prendre en compte l'imprédictibilité à l'horizon du mot pour le français ne fait que moyenner l'utilisation des autres catégories. En revanche, pour l'anglais, l'ajout de cette information aboutit à accorder plus d'importance à l'imprédictibilité à l'horizon de la syllabe également. Ainsi, d'après les modèles, la prosodie du locuteur français utilisé pourrait se situer d'avantage à l'horizon de la syllabe. Ceci serait cohérent avec le constat effectué précédemment concernant l'information de prédictibilité à l'horizon de la syllabe.

Ainsi, le système HTS considère l'information de prédictibilité comme information importante pour l'ensemble des deux langues. Toutefois, l'accent semble plutôt mis sur l'horizon de la syllabe pour le corpus français et sur l'horizon du mot pour le corpus anglais.

5 Conclusion

Dans cet article nous avons intégré un nouveau type de descripteurs linguistiques pour l'anglais et le français. Nous avons comparé l'influence de ces descripteurs sur la modélisation effectuée par le système HTS. Nous avons pu constater que cette information était considérée comme importante pour les deux corpus par le système HTS.

Toutefois, il reste difficile d'évaluer objectivement et subjectivement l'apport de tels descripteurs. En effet, les méthodes d'évaluation subjective permettant d'évaluer la qualité de modélisation des paramètres prosodiques, utilisés dans la synthèse de la parole, indépendamment des paramètres spectraux restent à déterminer.

6 Remerciements

La recherche présentée ici à été financé par la German Research Foundation (DFG) via le projet SFB 1102 « Information Density and Linguistic Encoding » à l'université de la Sarre. Nous souhaitons enfin remercier Marina Oberwegner pour la correction manuelle de la segmentation.

Références

BAUMANN T. & SCHLANGEN D. (2012). INPRO_iSS : a component for just-in-time incremental speech synthesis. In *Proceedings of the ACL 2012 System Demonstrations*, p. 103–108 : Association for Computational Linguistics.

HALE J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, p. 1–8 : Association for Computational Linguistics.

JAEGER T. F. (2010). Redundancy and reduction : speakers manage syntactic information density. *Cognitive Psychology*, **61**, 23–62.

KAWAHARA H., MASUDA-KATSUSE I. & DE CHEVEIGNÉ A. (1999). Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based F0 extraction : Possible role of a repetitive structure in sounds. **27**, 187–207.

KING S. & KARAISKOS V. (2013). The Blizzard Challenge 2013.

LE MAGUER S., BARBOT N. & BOEFFARD O. (2013). Evaluation of contextual descriptors for hmm-based speech synthesis in french. In *Proceedings of the Speech Synthesis Workshop (SSW)*, Barcelona (Spain).

LE MAGUER S., MÖBIUS B. & STEINER I. (2016). Toward the use of information density based descriptive features in hmm based speech synthesis. In *Proceedings of Speech Prosody*. to appear.

OBIN N., RODET X. & LACHERET A. (2010). Hmm-based prosodic structure model using rich linguistic context. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, p. 1133–1136.

POUGET M., HUEBER T., BAILLY G. & BAUMANN T. (2015). HMM training strategy for incremental speech synthesis. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*.

SCHRÖDER M. & TROUVAIN J. (2003). The German text-to-speech synthesis system MARY : A tool for research, development and teaching. *International Journal of Speech Technology*, **6**, 365–377.

SHANNON C. (1948). *A mathematical theory of distribution*, volume 27.

SMITH N. J. & LEVY R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, **128**(3), 302–319.

TOKUDA K., ZEN H. & BLACK A. W. (2002). An HMM-based speech synthesis system applied to English. In *Proceedings of the Speech Synthesis Workshop (SSW)*.

WANG P., QIAN Y., SOONG F. K., HE L. & ZHAO H. (2015). Word embedding for recurrent neural network based TTS synthesis. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, p. 4879–4883.

ZEN H., SENIOR A. & SCHUSTER M. (2013). Statistical parametric speech synthesis using deep neural networks. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, p. 7962–7966.

ZEN H. & TODA T. (2005). An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005. In *Proceedings of the European Conference on Speech Communicationand Technology (Eurospeech)*.

Utilisation des représentations continues des mots et des paramètres prosodiques pour la détection d'erreurs dans les transcriptions automatiques de la parole

Sahar Ghannay [1] Yannick Estève [1] Nathalie Camelin [1] Camille Dutrey [2, 3]
Fabián Santiago [2] Martine Adda-Decker [2, 4]
(1) Laboratoire d'Informatique de l'Université du Maine (LIUM), Avenue Laennec, Le Mans, France
(2) Laboratoire de Phonétique et Phonologie (LPP), 19 rue des Bernardins, Paris, France
(3) Laboratoire National de Métrologie et d'Essais (LNE), 29 avenue Roger Hennequin, Trappes, France
(4) Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur (LIMSI), Rue John Von Neumann, Orsay, France
{prenom.nom}@univ-lemans.fr[1] , {prenom.nom}@univ-paris3.fr[2]

RÉSUMÉ

Récemment, l'utilisation des représentations continues de mots a connu beaucoup de succès dans plusieurs tâches de traitement du langage naturel. Dans cet article, nous proposons d'étudier leur utilisation dans une architecture neuronale pour la tâche de détection des erreurs au sein de transcriptions automatiques de la parole. Nous avons également expérimenté et évalué l'utilisation de paramètres prosodiques en suppléments des paramètres classiques (lexicaux, syntaxiques, ...). La principale contribution de cet article porte sur la combinaison de différentes représentations continues de mots : plusieurs approches de combinaison sont proposées et évaluées afin de tirer profit de leurs complémentarités. Les expériences sont effectuées sur des transcriptions automatiques du corpus ETAPE générées par le système de reconnaissance automatique du LIUM. Les résultats obtenus sont meilleurs que ceux d'un système état de l'art basé sur les champs aléatoires conditionnels. Pour terminer, nous montrons que la mesure de confiance produite est particulièrement bien calibrée selon une évaluation en terme d'Entropie Croisée Normalisée (NCE).

ABSTRACT

Combining continuous word representation and prosodic features for ASR error detection

Recent advances in continuous word representation have been successfully used in several natural language processing tasks. This paper focuses on error prediction in Automatic Speech Recognition (ASR) outputs and proposes to investigate the use of continuous word representation (word embeddings) within a neural network architecture. The main contribution of this paper is about word embeddings combination : several combination approaches are proposed in order to take advantage of their complementarity. The use of prosodic features, in addition to classical syntactic ones, is evaluated. Experiments are made on automatic transcriptions generated by the LIUM ASR system applied on the ETAPE corpus. They show that the proposed neural architecture, using an effective continuous word representation combination and prosodic features as additional features, outperforms significantly state-of-the-art approach based on the use of Conditional Random Fields. Last, the proposed system produces a well calibrated confidence measure, evaluated in terms of NCE.

MOTS-CLÉS : Détection des erreurs de SRAP, réseau de neurones, représentation continue de mot, paramètres prosodiques.

KEYWORDS: ASR error detection, neural networks, word embeddings, prosodic features.

1 Introduction

Les avancées scientifiques récentes dans le domaine du traitement de la parole ainsi que la disponibilité d'une puissance de calcul croissante, ont conduit à l'obtention de performances très intéressantes d'un point de vue applicatif dans le domaine de la reconnaissance de la parole (SRAP). Cependant, malgré ces performances, les SRAPs génèrent encore des erreurs. Cela s'explique par leur sensibilité aux diverses variabilités liées à l' environnement acoustique, au locuteur, au style de langage, à la thématique du discours, *etc*). Ces erreurs présentent un obstacle à l'exploitation des transcriptions automatiques, par exemple pour certains traitements automatiques tels que l'extraction d'information, la traduction de la parole, la compréhension de la parole, *etc*.

L'exploitation efficace des transcriptions automatiques reste un but qui peut être atteint si l'on est capable de détecter et/ou de corriger des erreurs contenues dans les transcriptions automatiques. Cependant, la détection d'erreurs n'est pas une tâche facile, du fait qu'il existe plusieurs types d'erreurs. Ces erreurs peuvent aller de la simple substitution d'un mot par un homophone à l'insertion d'un mot non pertinent pour la compréhension globale de la séquence de mots. Elles peuvent aussi se répercuter sur les mots voisins et entraîner une séquence de mots erronés.

Dans cet article, nous proposons une architecture neuronale pour la détection d'erreurs. Nous nous intéressons à l'utilisation des représentations continues des mots [1], qui ont été introduites à l'origine pour la construction de modèles de langages neuronaux (Schwenk *et al.*, 2006). Ces représentations ont montré leurs impacts positifs dans de nombreuses tâches de traitement automatique du langage naturel (NLP) telles que : l'étiquetage morpho-syntaxique, le regroupement en syntagmes, la reconnaissance d'entités nommées ou encore l'étiquetage de rôles sémantiques (Turian *et al.*, 2010; Collobert *et al.*, 2011).

Dans le cadre de la tâche de détection d'erreurs de SRAP, nous étudions l'utilisation de plusieurs types de word embeddings provenant de différentes implémentations disponibles : `word2vec` (Mikolov *et al.*, 2013), `GloVe` (Pennington *et al.*, 2014) et une variante des embeddings de Collobert et Weston (Turian *et al.*, 2010). Afin de bénéficier de leurs potentielles complémentarités, nous proposons différentes approches de combinaison.

Le meilleur embedding obtenu par combinaison est utilisé pour représenter un mot reconnu, ainsi que les probabilités *a posteriori* de ce mot, des paramètres lexicaux, syntaxiques et prosodiques, qui sont intégrés dans une architecture neuronale pour une détection efficace des erreurs. Bien que les paramètres prosodiques aient déjà été utilisés dans le cadre de la détection des erreurs au niveau de l'énoncé (utterance), leur utilisation pour la détection des erreurs au niveau du mot a été moins étudiée.

Enfin, nous nous intéresserons à l'évaluation des mesures de confiance produites par le système neuronal.

Cet article est organisé de la manière suivante : la section 2 présente les travaux liés à la tâche de détection des erreurs, l'intégration des embeddings et des paramètres prosodiques pour cette tâche. La section 3 détaille les différents types d'embeddings ainsi que les approches proposeés pour les combiner. La section 4 décrit le système de détection d'erreurs proposé. Le protocole expérimental et les résultats sont décrits dans la section 5, juste avant la conclusion (Section 6).

1. par la suite on utilisera la terminologie anglaise *word embeddings*

2 Travaux connexes

Depuis deux décennies, de nombreuses études se focalisent sur la détection des erreurs de SRAP.

Plusieurs approches sont fondées sur l'utilisation des champs aléatoires conditionnels (CRF). Dans (Parada *et al.*, 2010), les auteurs se sont intéressés à la détection des régions d'erreurs générées par les mots hors vocabulaires en prenant en compte des informations contextuelles des régions. Une approche similaire pour d'autres types d'erreurs a été présentée dans (Béchet & Favre, 2013). Celle-ci est basée sur un étiqueteur de séquence à base de CRF utilisant des paramètres issus de systèmes de transcription, des paramètres lexicaux et syntaxiques.

L'approche la plus récente utilise un réseau de neurones qui intègre plusieurs sources d'information afin de détecter si un mot est correct ou erroné (Yik-Cheung *et al.*, 2014). On peut notamment citer des paramètres extraits à partir du modèle de langage basé sur les réseaux de neurones récurrents, des réseaux de confusion. D'autres paramètres proviennent également de la complémentarité de deux SRAPs.

Les embeddings constituent une projection des mots du vocabulaire dans un espace de faible dimension. Ils sont utilisés avec succès, comme paramètres supplémentaires, dans plusieurs tâches NLP (Turian *et al.*, 2010; Collobert *et al.*, 2011). Les auteurs dans (Turian *et al.*, 2010) ont évalué différents types de word embeddings ainsi que leur combinaison par simple concaténation pour la tâche de reconnaissance d'entités nommées et le regroupement en syntagmes.

Notre utilisation des paramètres prosodiques est motivée par de précédents travaux, notamment (Stoyanchev *et al.*, 2012) et (Goldwater *et al.*, 2010). Les premiers ont montré que la combinaison des paramètres prosodiques et syntaxiques est utile pour localiser les mots mal reconnus dans un tour de parole. Les seconds ont découvert que les mots mal reconnus ont des valeurs prosodiques extrêmes.

Dans cette étude, nous proposons d'intégrer la meilleure combinaison des embeddings avec d'autres paramètres supplémentaires dans une architecture neuronale conçue pour la détection des erreurs de SRAP.

3 Représentations continues (*embeddings*) de mot

Différentes approches ont été introduites pour calculer les embeddings de mots à travers les réseaux de neurones.

Dans le cadre de la détection des erreurs, nous avons besoin de capturer des informations syntaxiques afin de les utiliser pour analyser les séquences de mots reconnus, mais nous avons aussi besoin de capturer des informations sémantiques pour mesurer la pertinence des co-occurrences de mots dans la même hypothèse. Nous avons utilisé et évalué, trois types d'embeddings provenant de différentes implémentations disponibles (plus de détails dans (Ghannay *et al.*, 2015a)).

3.1 Description des embeddings de mots

Trois types d'embeddings, à 100 dimensions chacun, ont été calculés à partir d'un vaste corpus textuel composé d'environ 2 milliards de mots. Ce corpus a été construit à partir des articles du journal français *Le Monde*, le corpus *Gigaword*, les articles fournis par *Google News* et les transcriptions manuelles d'environ 400 heures d'émissions françaises.

Ces embeddings sont détaillés dans ce qui suit :

tur : embeddings de Collobert et Weston (Turian *et al.*, 2010), revisités par Joseph Turian. Ils

sont basés sur l'existence ou non de n-grammes dans le corpus d'apprentissage.

w2v-CBOW : calculés avec la boite à outils *word2vec*. Ils sont estimés avec l'approche sac de mots continus (*Continuous bag-of-words (CBOW)*).

GloVe : basés sur l'analyse des co-occurrences des mots dans une fenêtre (Pennington *et al.*, 2014).

3.2 Combinaison des embeddings

Afin de tirer profit de la complémentarité des embeddings décrits ci-dessus, nous avons proposé de les combiner en utilisant plusieurs approches détaillées ci-après.

3.2.1 Concaténation simple

La première approche est inspirée de celle proposée par (Turian *et al.*, 2010). Elle consiste à concaténer les embeddings selon cet ordre : *GloVe*, *tur* et *w2v-CBOW*, (nommé ***GTW***). Chaque mot est ainsi représenté par un vecteur de taille 300.

3.2.2 Analyse en Composantes Principales (ACP)

Cette deuxième approche consiste à transformer des variables corrélées entre elles en nouvelles variables décorrélées les unes des autres (appelées composantes principales ou axes principaux). Les premiers axes portent plus d'informations que les derniers (en terme de dispersion de données). L'ACP est appliquée au vecteur *GTW* (celui obtenu par simple concaténation). Elle est calculée en utilisant la matrice de corrélation pour obtenir le nouveau système vectoriel. Ce système est projeté ensuite dans une nouvelle base. Nous considérons par la suite uniquement les 200 premières composantes du vecteur projeté (nommé ***GTW-PCA-200***).

3.2.3 Auto-encodeurs

La troisième approche se base sur l'utilisation d'auto-encodeurs ordinaire (*O*) et de débruitage (*D*). Ces auto-encodeurs sont composés d'une couche cachée contenant 200 (*GTW-200*) unités cachées chacune. Ils prennent en entrée le vecteur *GTW* et génèrent en sortie un vecteur de 300 unités. Pour chaque mot, le vecteur des valeurs numériques produites par la couche cachée sera utilisé comme embeddings combiné (nommés ***GTW-D/GTW-O***).

La différence entre l'auto-encodeur ordinaire et de débruitage vient de la notion de corruption aléatoire des entrées au cours de l'apprentissage dans le dernier cas. La corruption consiste à initialiser une partie des données à zéro. Cette corruption rend l'auto-encodeur plus généraliste en découvrant des paramètres plus robustes qu'un auto-encodeur ordinaire. À notre connaissance, les auto-encodeurs ne sont pas utilisés pour la combinaison, mais juste utilisés pour apprendre une représentation compressée pour un ensemble de données, généralement dans le but de réduire le nombre de dimensions.

4 Système de détection d'erreurs

Le système de détection d'erreurs doit attribuer une étiquette *correcte* ou *erreur* à chaque mot en se basant sur un ensemble de paramètres. Cette attribution est faite en analysant chaque mot dans son contexte.

4.1 Paramètres utilisés

Dans cette section, nous décrivons les paramètres recueillis pour chaque mot et comment ceux-ci sont extraits. Certains de ces paramètres sont identiques à ceux présentés dans (Béchet & Favre, 2013).

Chaque mot est représenté par un vecteur composé des paramètres suivants :

Mesures de confiance du SRAP : il s'agit des probabilités *a posteriori* (PAP) générées par le SRAP. La PAP est calculée à partir du réseau de confusion qui est approximée par la somme des probabilités *a posteriori* de toutes les transitions passant par ce mot et qui sont en concurrence avec lui.

Paramètres lexicaux : ils se composent de la 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.

Paramètres syntaxiques : ils se composent de l'étiquette syntaxique (POS) et du gouverneur du mot courant ainsi que des liens de dépendances entre le mot courant et son gouverneur. [2]

Paramètres prosodiques : il s'agit de deux ensembles de paramètres. Les premiers sont extraits à partir de l'alignement forcé des transcriptions avec le signal audio : nombre de phonèmes, durée moyenne des phonèmes et durée de la pause précédent le mot. Le dernier paramètre prosodique correspond à la fréquence F0 [3]. Ces paramètres sont détaillés dans (Ghannay *et al.*, 2015b).

Mot : il s'agit de la représentation orthographique du mot dans l'étiqueteur de séquence à base de CRF et de son embedding dans le système neuronal.

4.2 Architecture

Nous utilisons une architecture neuronale basée sur une stratégie multi-flux pour l'apprentissage du réseau de neurones, nommée Perceptron Multicouche MultiStream (*MLP-MS*). Une description détaillée de cette architecture est présentée dans (Ghannay *et al.*, 2015a). Cette architecture illustrée dans la figure 1 est utilisée afin de mieux intégrer des informations contextuelles à partir des mots voisins.

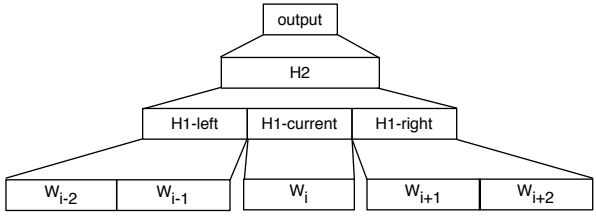


FIGURE 1: L'architecture MLP-MS pour la détection des erreurs de SRAP.

5 Expériences et résultats

5.1 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*. Il s'agit d'un système de transcription 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).

2. Ces paramètres sont fournis par la boite à outils MACAON http ://macaon.lif.univ-mrs.fr
3. La F0 est obtenue en analysant le signal avec la boite à outils Praat [4]

Les transcriptions automatiques ont été alignées avec les transcriptions de référence en utilisant l'outil *sclite* [5]. 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: Description des données expérimentales.

5.2 Résultats expérimentaux

Cette section présente les résultats expérimentaux de notre système de détection d'erreurs *MLP-MS* et les compare à un système état de l'art basé sur les CRFs implémenté avec *Wapiti* [6]. Les résultats sont évalués en termes de rappel (R), précision (P) et F-mesure (F) pour la détection de mots erronés et le taux d'erreur de classification globale (CER). La mesure de confiance calibrée produite par les systèmes de détection d'erreurs est évaluée en terme d'Entropie Croisée Normalisée (NCE). Enfin, la significativité de nos résultats est mesurée en utilisant un intervalle de confiance à 95 %.

Afin de mesurer plus particulièrement l'apport des paramètres prosodiques, l'ensemble des paramètres présentés en section 4.1 exceptés les paramètres prosodiques sont utilisés en section 5.2.1 puis la section 5.2.2 présente les résultats lorsque tous les paramètres sont utilisés.

5.2.1 Performance des différents embeddings de mots dans le système neuronal

Un ensemble d'expériences est effectué afin d'évaluer l'impact des différents types d'embeddings ainsi que celui de leurs combinaisons. Les systèmes de détection d'erreurs (MLP-MS et CRF) sont entrainés sur le corpus d'apprentissage *Train* et optimisés sur le corpus de développement *Dev*. Les résultats expérimentaux présentés dans le tableau 2 montrent que notre proposition de combiner les embeddings est utile et améliore significativement les résultats en termes de *CER* par rapport à l'utilisation d'embeddings non-combinés. Les meilleures combinaisons *GTW-D200* et *GTW-O200* conduisent à une réduction du CER comprise entre 5 % et 5,3 % par rapport à l'approche CRF. Ces deux embeddings combinés sont utilisés dans la suite des expériences.

		Label erreur			Globale
Approches	**Représentation**	**P**	**R**	**F**	**CER**
Neuronale	glove	67.80	53.23	59.64	10.60
	tur	70.63	48.61	57.58	10.54
	w2v	72.25	46.65	56.69	10.49
	GTW 300	69.68	52.23	59.71	10.38
	GTW-PCA200	71.82	47.37	57.09	10.48
	GTW-O200	71.78	54.35	61.86	**9.86**
	GTW-D200	69.61	58.24	63.42	**9.89**
CRF	discrète	69.67	51.89	59.48	10.41

TABLE 2: Comparaison de différents types d'embeddings dans *MLP-MS* sur le corpus *Dev*

5. http ://www.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm
6. http ://wapiti.limsi.fr

Le tableau 3 compare les performances des embeddings *GTW-D200* et *GTW-O200* à celles du système CRF sur le corpus de test. Ces représentations réalisent respectivement 6.14% et 7.84% de réduction de CER par rapport aux CRF.

		Label erreur			Globale	
Corpus	**Approche**	**P**	**R**	**F**	**CER**	**95% intervalle de confiance**
Test	CRF	68.34	49.66	57.52	8.79	[8.55 ; 9.04]
	GTW-O200	**71.00**	54.76	61.83	**8.10**	[7.87 ; 8.33]
	GTW-D200	68.23	**58.35**	**62.90**	8.25	[8.01 ; 8.49]

TABLE 3: Performance des meilleurs embeddings sur le corpus *Test*.

5.2.2 Performance des paramètres prosodiques

Le tableau 4 présente l'impact des paramètres prosodiques sur les résultats présentés ci-dessus. L'ajout de ces paramètres conduit à une réduction du CER par rapport aux résultats des tableaux 2 et 3. De plus, nos systèmes *GTW-D200+PROS* et *GTW-O200+PROS* obtiennent des améliorations significatives par rapport aux CRF respectivement de 8,65 % et 9,45 % de réduction en CER.

		Label erreur			Globale	
Corpus	**Approche**	**P**	**R**	**F**	**CER**	**95% confiance interval**
Dev	CRF+PROS	69.89	52.86	60.20	10.29	[10.03 ; 10.56]
	GTW-O200+PROS	69.76	60.50	64.80	9.67	[9.40 ; 9.93]
	GTW-D200+PROS	**70.40**	**60.57**	**65.11**	**9.55**	[9.30 ; 9.81]
Test	CRF+PROS	68.95	51.82	59.17	8.57	[8.33 ; 8.81]
	GTW-O200+PROS	**69.01**	**60.95**	**64.73**	**7.96**	[7.73 ; 8.20]
	GTW-D200+PROS	68.68	60.65	64.42	8.03	[7.80 ; 8.27]

TABLE 4: Performance des paramètres prosodiques sur les corpus *Dev* et *Test*.

5.2.3 Calibration des mesures de confiance

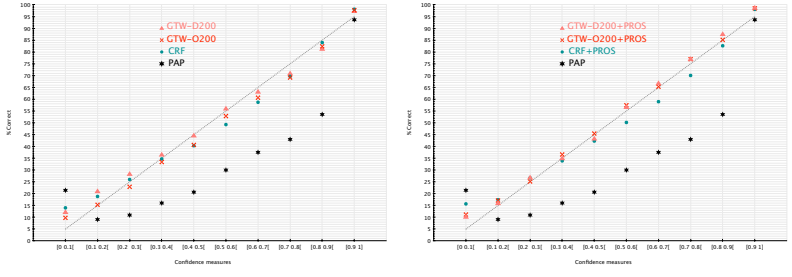
Dans une volonté d'améliorer la qualité des mesures de confiance, le problème de leur calibration a été abordé dans (Yu *et al.*, 2011). Cette étape de post-traitement est considérée comme une technique d'adaptation spéciale appliquée à la mesure de confiance afin de prendre des décisions optimales. L'utilisation de réseaux de neurones artificiels est l'une des méthodes proposées pour cette étape de post-traitement.

D'après les scores NCE présentés dans le tableau 5, nous démontrons la validité de notre approche pour produire une mesure de confiance bien calibrée, tandis que la probabilité *a posteriori* fournie par le système LIUM SRAP n'est pas calibrée. Le système CRF produit également une mesure de confiance bien calibrée. En outre, l'utilisation des caractéristiques prosodiques améliore les scores des NCE pour tous les systèmes.

Comme le montre la figure 2, les probabilités dérivées de nos systèmes neuronaux et CRF correspondent à la probabilité des mots corrects. En outre, les courbes sont bien alignées avec la diagonale, en particulier pour nos systèmes neuronaux avec des paramètres prosodiques.

Name	PAP	proba softmax GTW-D200	proba softmax GTW-O200	CRF
sans paramètres prosodiques				
Dev	-0.064	0.425	0.443	**0.445**
Tes	-0.044	0.448	**0.461**	0.457
avec paramètres prosodiques				
Dev	-0.064	0.461	**0.463**	0.449
Test	-0.044	0.471	**0.477**	0.463

TABLE 5: score NCE pour la PAP et les mesures de confiances issues des systèmes MLP-MS et CRF.



(a) (b)

FIGURE 2: Pourcentage de mots corrects en fonction des scores de la PAP et des mesures de confiances issues des systèmes MLP-MS et CRF sans (a) et avec paramètres prosodiques (b).

6 Conclusion

Dans cet article, nous avons évalué l'intégration de différentes représentations continues de mot (embeddings) sur la tâche de détection d'erreurs de SRAP. La partie expérimentale, effectuée sur le corpus ETAPE, a montré la validité de notre approche pour la détection des erreurs. Nous avons notamment proposé différents types d'embeddings simples et combinés et montré le gain obtenu par l'utilisation des embeddings combinés. De plus, nous avons prouvé l'apport significatif de l'utilisation des paramètres prosodiques, en plus de ceux syntaxiques et lexicaux classiques. En outre, les résultats obtenus sont meilleurs que ceux d'un système état de l'art basé sur les champs aléatoires conditionnels. Pour terminer, nous nous sommes attachés à démontrer que les mesures de confiances produites par notre système sont bien calibrées.

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 localisation for spoken recovery strategy. In *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference*.

COLLOBERT R., WESTON J., BOTTOU L., KARLEN M., KAVUKCUOGLU K. & KUKSA P. (2011). Natural Language Processing (Almost) from Scratch. volume 12, p. 2493–2537 : JMLR.

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.

GHANNAY S., ESTÈVE Y. & CAMELIN N. (2015a). 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., DUTREY C., SANTIAGO F. & ADDA-DECKER M. (2015b). Combining continuous word representation and prosodic features for asr error prediction. In A.-H. DEDIU, C. MARTÍN-VIDE & K. VICSI, Eds., *Statistical Language and Speech Processing*, volume 9449 of *Lecture Notes in Computer Science*, p. 84–95. Springer International Publishing.

GOLDWATER S., JURAFSKY D. & MANNING C. D. (2010). Which words are hard to recognize ? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, p. 181–200.

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)*.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

PARADA C., DREDZE M., FILIMONOV D. & JELINEK F. (2010). Contextual information improves OOV detection in speech. In *Human Language Technologies : Proceddings of the North American Chapter of the Association for Computational Linguistics (NAACL'10)*.

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.

SCHWENK H., DCHELOTTE D. & GAUVAIN J.-L. (2006). Continuous space language models for statistical machine translation. In *Proceedings of COLING/ACL*, COLING-ACL '06, p. 723–730, Stroudsburg, PA, USA : Association for Computational Linguistics.

STOYANCHEV S., SALLETMAYR P., YANG J. & HIRSCHBERG J. (2012). Localized detection of speech recognition errors. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, p. 25–30.

TURIAN J., RATINOV L. & BENGIO Y. (2010). Word representations : A simple and general method for semisupervised learning. p. 384–394.

YIK-CHEUNG T., LEI Y., ZHENG J. & WANG W. (2014). ASR error detection using recurrent neural network language model and complementary ASR. In *Proceddings of Acoustics, Speech and Signal Processing (ICASSP 2014)*, p. 2312–2316.

YU D., LI J. & DENG L. (2011). Calibration of confidence measures in speech recognition. In *IEEE Transactions on Audio, Speech, and Language Processing*, volume 19, p. 2461–2473.

Variabilité des syllabes réalisées par des apprenants de l'anglais

Nicolas Ballier[1] Philippe Martin[2] Maelle Amand[3]
[1, 3] Université Paris-Diderot, EA 3967 – CLILLAC-ARP, France
[2] Université Paris-Diderot, UMR 7110-Laboratoire de Linguistique Formelle, France
nicolas.ballier@univ-paris-diderot.fr,
philippe.martin@linguist.univ-paris-diderot.fr,
maelle.amand@gmail.com

RÉSUMÉ

Cette contribution analyse la segmentation syllabique des francophones du corpus d'apprenant d'anglais ANGLISH (Tortel 2009). A partir d'une méthode d'alignement par alignement forcé, on montre la pertinence d'une analyse de l'interlangue fondée sur la comparaison des durées des syllabes. La comparaison des réalisations est ici centrée sur une typologie des syllabes fondée sur des propriétés distributionnelles, accentuelles et où l'interlangue tient sa place (risques d'isosyllabicité les plus manifestes pour les réalisations des francophones). La variabilité des réalisations des syllabes est appréciée en fonction des propriétés positionnelles, accentuelles et structurelles des syllabes.

L'étude démontre l'intérêt d'une approche fonctionnelle des syllabes, plus pertinente que les intervalles interconsonantiques et intervocaliques inspirés de Ramus et al. (1999) pour la discrimination du niveau des locuteurs.

ABSTRACT

Analysing syllable variability in a French learner corpus of English.
This paper suggests an alternative method to classify native and non-native rhythmic realisations. Learner phonetic output has been automatically aligned on a native template of English syllables. Syllables have been classified according to positional, distributional and accentual properties. These syllable types differ significantly in their realisations between native and non-native speakers of English.

MOTS-CLÉS : syllabation, interphonologie français/anglais, alignement forcé, durée.

KEYWORDS: syllables, French-English interphonology, forced alignment, duration, templatic transfers.

1. Introduction

L'importance de la place du rythme dans l'interphonologie français/anglais ne fait pas débat (Adams 1973 ; Tortel 2009; Herment 2015). Toute la difficulté réside dans la prise en compte de la mesure idoine du rythme. Nous proposons de sortir de l'inventivité statistique que l'on peut associer aux mesures du rythme (Arvanati 2012; Loukina et al. 2013 ; Gut 2014) et de revenir aux constituants prosodiques (syllabe accentuée, syllabe réduite, pied). Sortir de l'impasse des métriques, dominantes depuis Ramus et al. (1999), permet de restituer les constituants de la hiérarchie prosodique, et de montrer la pertinence de la syllabe accentuée et du pied pour l'analyse de l'anglais. En l'espèce, on peut décrire l'interlangue prosodique des francophones parlant anglais, et aborder la question de l'isosyllabicité et de la structuration accentuelle dans le cadre d'une réflexion ciblée. On cherche à comparer la prosodie des locuteurs sur des séquences identiques de texte lu. Nous proposons de considérer la variabilité dans deux dimensions.

L'approche paradigmatique consistera ultérieurement à comparer les événements prosodiques produits sur un intervalle donné. L'approche syntagmatique compare les différences de durée sur des

empans comparables, ce qui pose la question de la discrétisation des intervalles pertinents à mesurer pour construire les comparaisons. L'approche suivie par (Ballier et al, 2014) avec des outils comme Prosody Pro (Xu 2013) permet de normaliser la durée des réalisations en ne comparant que le matériau syllabique (les réalisations prosodiques, syllabe par syllabe). En proposant un compromis des réalisations prosodiques (visualisation de la moyenne de F0 sur l'intervalle de durée de la syllabe), elle occulte toute la complexité et induit une représentation des variations de la F0 fictive, décorrélée des points cibles mais figurée à partir de la médiane de l'intervalle syllabique. En résumé, ce type d'approche normalise la durée et réintroduit une comparaison fondée sur un constituant prosodique, mais occulte les variations rythmiques et simplifie les événements prosodiques.

S'agissant du rythme, (Klatt 1987)[1] posait déjà la problématique de l'empan de la mesure des différentiels temporels. Notre contribution, après plus de dix-sept ans de recherche depuis l'article fondateur de Ramus et al. (1999) et au moins autant de mesures différentes du rythme (Loukina et al. 2013 pour une synthèse de quinze métriques), est de faire porter les mesures temporelles sur des unités linguistiques plus fondées que les intervalles interconsonantiques ou intervocaliques pour l'étude de la L2 (Ballier 2016) et de proposer un différentiel natif/non-natif. En particulier, nous proposons une typologie fonctionnelle des syllabes. On cherche à établir les zones de stabilité des réalisations des natifs dans la parole lue susceptibles de proposer des patrons de comparaison, des « modèles », réalisations où nos postulons que la grammaire l'emporte sur la variation phonostylistique des idiosyncrasies du locuteur.

Deux procédés sont possibles. Le premier consiste à travailler au niveau de l'alignement du phone et permet éventuellement de comparer la réalisation des francophones et des anglophones. Dans le cas d'un alignement au niveau du phone, on peut espérer détecter de manière automatique les écarts de réalisation des phones, on s'attachera plus particulièrement à la non-réduction vocalique, aux voyelles épenthétiques et pour la réalisation consonantique des diphtongues. Dans le deuxième cas qui nous intéresse, l'analyse porte sur le découpage de portions du signal, on cherche ainsi à établir le degré de comparabilité des séquences syllabiques. Dans cette entreprise, on commence par établir un patron de réalisation des syllabes anglophones. On a pris soin de délimiter l'intégralité des consonnes syllabiques du corpus. À cet effet, le découpage en syllabes des réalisations des anglophones sert de patron pour l'alignement des réalisations des francophones. La première étape consiste en un repérage des séquences pertinentes pour l'analyse. Dans le logiciel WinPitch LTL, la configuration passe par une paramétrisation des couleurs qui permettent de repérer, selon le choix de l'utilisateur, des segments temporels de la fréquence fondamentale, l'intensité ou le spectrogramme.

2. Interphonologie du rythme

La problématique la plus générale est celle qui porte sur la comparabilité des réalisations au niveau prosodique, que l'on aborde ici à partir de la question de la réalisation des syllabes. Dans cette contribution, on considère que le rythme est la différence essentielle entre francophones et anglophones, de sorte que la mesure de différentiel temporel inter-syllabique est suffisante. A la grande différence de l'impasse des métriques du rythme, qui portent sur des intervalles interconsonantiques ou intervocaliques, notre analyse repose sur des entités linguistiques attestées, à savoir les syllabes. On s'intéresse donc dans cette perspective à mesurer le différentiel temporel de réalisation des syllabes entre productions des francophones et productions des anglophones. En ce sens, la notation porte une comparaison des réalisations et des différentiels entre syllabes accentuées et syllabes inaccentuées. Avantage considérable, cette méthode permet ultérieurement de rétablir l'unité prosodique la plus idoine : le pied (Abercrombie 1965; Roach 1983 et toute la tradition de la

[1] "one of the unsolved problems in the development of rule systems for speech timing is the size of the unit (segment, onset/rhyme, syllable, word) best employed to capture various timing phenomena" (Klatt 1987 : 760).

phonologie britannique). Afin d'affiner l'analyse des différences entre syllabes accentuées et inaccentuées (Gut 2003), nous complexifions la typologie des syllabes analysées.

Nous ne traitons pas ici des « ré-alignements tonaux » (Mennen 2015), en partie parce que la synchronisation des *pitch targets* et des syllabes reste problématique (Xu & Xu 2005 : 193). Nous ne limitons pas l'analyse aux polysyllabiques et comparons aussi les (nombreuses) syllabes individuelles des monosyllabes.

Dauer (1983) et Crystal (1996) expliquent l'importance des schémas syllabiques dans l'analyse du rythme. Nous faisons l'hypothèse que certains types de syllabes devraient constituer des zones de stabilité dans les réalisations des natifs, d'autres étant davantage sujettes à la variation phonostylistique. Certains types de syllabes, en particulier les voyelles réduites, les formes faibles et les réalisations syllabiques sont plus spécifiques au système phonologique des anglophones et devraient davantage refléter les différences entre groupes d'apprenants FR1 et FR2 et réalisations des anglophones.

3. Corpus et méthodes

Pour reprendre le distinguo de Mennen (2015), on s'intéressera aux différences réalisationnelles plus que sémantiques des événements prosodiques. La comparaison sera centrée sur l'analyse du rythme.

3.1 Description du corpus

Il s'agit d'une partie des passages lus du corpus ANGLISH (Tortel & Hirst 2010). Quatre passages du corpus EuRom01 ont été lus par vingt anglophones natifs, vingt francophones de niveau intermédiaire (FR1) et vingt francophones en troisième année (FR2) de Licence d'anglais (Tortel 2009 ; Tortel 2013). Les anglophones ont un certain niveau de connaissances de français mais leur durée de résidence en France (un an en moyenne) n'a pas eu d'incidence sur leurs réalisations phonétiques de l'anglais.

3.2 Ontologie des syllabes-cibles annotées / typologie fonctionnelle des syllabes

Nous présentons d'abord la méthodologie d'analyse. La première expérience consiste à comparer les réalisations de 20 locuteurs anglophones natifs. Deux experts ont annoté les syllabes sur une partie du corpus (écart inter-annotateur; k=90 %) en établissant 99 « valeurs cibles » des 65 mots retenus à partir de la division syllabique du *Longman Pronouncing Dictionary*, qui privilégie le rattachement des consonnes en coda en cas de syllabe accentuée (Wells 1990). L'ontologie du découpage syllabique retenu prend en charge les types de syllabes suivants, récapitulés dans le Tableau 1, retenus en fonction de leur configuration (CV, CVC, CVCC), et des transferts attendus chez les francophones (réalisation d'une voyelle épenthétique en lieu de la voyelle réduite ou de la réalisation syllabique de la consonne [n] ou [l]) ainsi que leur statut accentuel dans la hiérarchie accentuelle. Dans cette analyse, la hiérarchie accentuelle a été simplifiée : quatre degrés sont pris en compte : syllabe sous accent principal, syllabe sous accent secondaire, inaccentuée non-réduite et inaccentuée réduite. Le corpus annoté ne comprenait pas d'accent secondaire au sens de Guierre (1979) et seulement deux composés.

Les propriétés positionnelles dans l'énoncé, distributionnelles dans la structure de syllabe ou le placement accentuel ne sont pas mutuellement exclusives. Pour les syllabes relevant de plusieurs cas, la hiérarchie suivante des critères a été suivie pour un codage unique de propriété : focus > position dans l'énoncé > nombre de syllabes > structure de syllabe.

Nous n'avons pas codé l'opposition des schwas pré- et post-toniques, même si les différences réalisationnelles sont importantes (Cruttenden 2015). La syllabe ouverte réduite (comme dans l'initiale de *forward*) ne reçoit pas ici de statut particulier. Nous avons retenu la distinction entre des [i] des *happY tensing* finaux et intertoniques comme dans *terrified*) mais pas opéré de distinction de timbre entre schwa et /ɪ/ pour les voyelles réduites. La logique qui a présidé au système d'annotation

est la pénalisation de l'isosyllabicité, que ce soit dans la maximalisation de la coda ou dans l'analyse des voyelles intertoniques. Pour la division après les rhotiques, on a suivi la syllabation du dictionnaire de Wells, indépendamment de la question de l'ambisyllabicité.

Comme cette annotation a pu le rappeler, établir un patron de réalisations attendues, mêmes pour des natifs, pose la question de la variabilité pour l'anglais et de la norme de référence pour la prononciation. Les deux problématiques sont donc alors : 1) Quelle est l'homogénéité des réalisations des anglophones (types de syllabes, proportion des durées) ? 2) Quelles sont les prévisions permises par l'alignement pour les réalisations des francophones (allongement et présomption de voyelles épenthétiques, proportion des durées selon les types de syllabes) ?

type de syllabe	caractérisation	interférences et transferts possibles
Focus	syllabe obligatoirement noyau de l'unité intonative (ex: *do* d'emphase)	réalisation de contour prosodique différente
Final	syllabe en fin d'unité intonative	risque de surallongement par transfert du français (Delais, Herment et al.)
MaxCoda	syllabe accentuée fermée par une ou plusieurs consonnes VC(C)(C) ex : /ˈ**brɪt** n/	possibilité de voyelle épenthétique, de rattachement de la coda à la syllabe suivante ([ˈbri ˈtoen])
RedFinCC	syllabe réduite en VC(C)	possibilité de voyelle épenthétique
Syllabiq	consonne syllabique, sommet de syllabe (n, l ou r) ex : /ˈbrɪt **n**/	possibilité de voyelle épenthétique [bri tœn]
Interton	schwa en position médiale	risque d'allongement
RedFin	syllabe réduite ouverte (en milieu de mot portant un accent)	allongement (risque de transfert d'accent démarcatif final du groupe accentuel)
RedCoda	syllabe réduite en coda CVC	
happY	syllabe inaccentuée ouverte (en finale de mot, happY /ˈhæp i/	risque d'allongement
preCV	syllabe pré-tonique ouverte (*a* dans *along*)	risque d'allongement vocalique
preCC	syllabe pré-tonique fermée (*in* dans *inspector*)	risque d'allongement, moindre que pour preCV.
monCV	monosyllabe en syllabe ouverte	risque de réalisation isosyllabique, allongements des voyelles inaccentuées
monCC	monosyllabe en syllabe fermée	risque de réalisation isosyllabique, allongements des voyelles inaccentuées
ff	forme faible : marqueur grammatical réalisé en voyelle réduite sauf en cas de focus (*of, the, a*)	risque d'isosyllabicité par allongement et réalisation de timbre vocalique autre que schwa (voyelle pleine).

TABLE 1 : Ontologie des syllabes annotées dans le corpus

3.3 Alignement des segments annotés

Nous avons segmenté les spectrogrammes des polysyllabiques à l'aide de la version de développement de WinPitch LTL (WinPitch 2016) et nous avons procédé à la comparaison des spectres et des intervalles temporels par alignement forcé (Bellman 1957). La méthode est relativement robuste. Avec un processeur Intel I7 à 2,1 Mhz de fréquence et 8 gigas de mémoire vive,

l'opération a pris 95s pour un fichier modèle de vingt-cinq secondes. Une optimisation de calcul évitant l'emploi des routines standard de C++ a permis d'obtenir un temps de calcul plus rapide. L'opération de comparaison se fait deux par deux et a été conduite à partir d'un fichier natif identifié comme particulièrement prototypique. Nous avons testé l'alignement obtenu à partir d'un deuxième fichier de natif et la différence ne s'est pas avérée significative.

3.4 Non-traitement des pauses

L'alignement forcé pose un problème lorsque la séquence des phones prononcés par les apprenants ne correspond pas à la séquence modèle. Cette caractéristique perturbe particulièrement l'alignement lors de la réalisation par les apprenants de pauses inexistantes dans le modèle. Les résultats peuvent être alors faussés et peuvent se révéler, par exemple par une durée excessive de tenue d'occlusive. L'algorithme que nous avons développé résout ce problème en ne tenant pas compte des pauses identifiées dans la recherche d'un chemin optimum dans la matrice d'alignement. L'avantage de cette méthode pour le traitement des corpus d'apprenants en parole lue est que l'on n'est pas dépendant des modèles de reconnaissance ou des dictionnaires utilisées, ce qui est le cas de SPPAS (Bigi 2012). L'erreur de segmentation est maximalement de 10 ms, correspondant au pas de la grille matricielle utilisée pour l'alignement. Il va de soi que d'autres méthodes simples suffiraient pour traiter l'anarchie des pauses dans le discours d'apprenants.

4 Résultats

Nous concentrons notre analyse sur les différents types de syllabes répertoriés parmi les 6.000 intervalles étudiés.

Comme le montre la figure 1, les moyennes des types de syllabes sont inégalement distinctives d'un groupe à un autre. Comme dans la thèse d'Anne Tortel, la supériorité de niveau des apprenants de niveau FR2 sur FR1 n'est pas probante. L'étalonnage des niveaux d'apprenants a été fait sur la base de la pratique des locuteurs plus que par un test initial de niveau type TOEFL. Ce redécoupage exact en syllabes permet néanmoins de faire la démonstration de la pertinence de l'aligneur pour un redécoupage en syllabes de tout le corpus. Les réalisations des francophones ne confirment pas toutes les prédictions : les syllabes réduites (ff) ne sont pas significativement différentes par la durée des réalisations des anglophones.

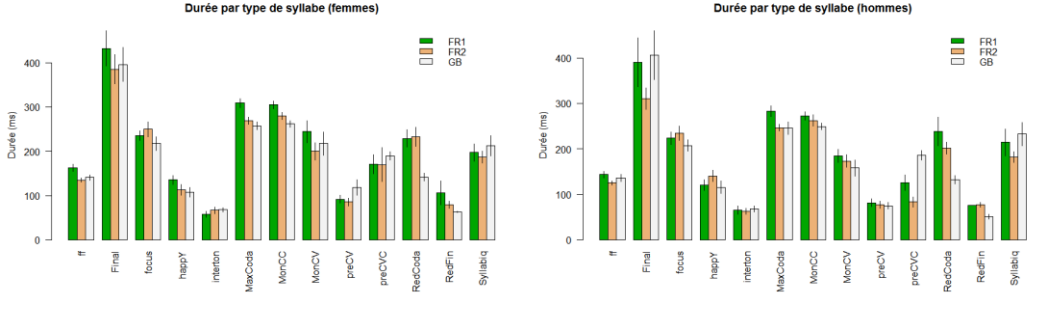


FIGURE 1 : Durées moyennes (ms) des types de syllabes pour les différents groupes

Grâce à cette étude préliminaire nous avons pu tester la stabilité de la réalisation de certains types de syllabes typologiquement éloignées du français (sous accent lexical, réduite, syllabique) et d'en analyser la variance. Les conditions n'étaient pas réunies pour procéder à des ANOVA (absence de normalité des données, $p < 0{,}01$). Le nombre de d'observations supérieur à 5.000 et des valeurs ex-aequo nous ont fait opter pour un test d'Anderson-Darling, qui a confirmé la nécessité d'avoir recours à un test non-paramétrique. Le test de Kruskal-Wallis nous a permis de comparer les médianes. Il n'y

a pas de différence de durée significative au niveau des médianes et des moyennes entre les FR2 et GB ($p > 0{,}01$), mais une différence significative entre GB et FR1 et FR1 et FR2 ($p < 0.01$). Par rapport aux GB, les FR1 sont en général plus lents de 23,7 ms (différence significative). Ce n'est pas le cas pour FR2, légèrement plus lents que les natifs, mais pas assez pour être en significativement différents. On peut repérer un effet de la variable genre, les hommes sont significativement plus rapides dans l'ensemble que les femmes.

Nous avons ensuite procédé à un test de comparaison multiple deux à deux pour les syllabes les plus significatives chez les anglophones. En figure 2, nous représentons les 10 paires de types de syllabes dont les durées moyennes sont les plus significativement différentes (valeurs absolues).

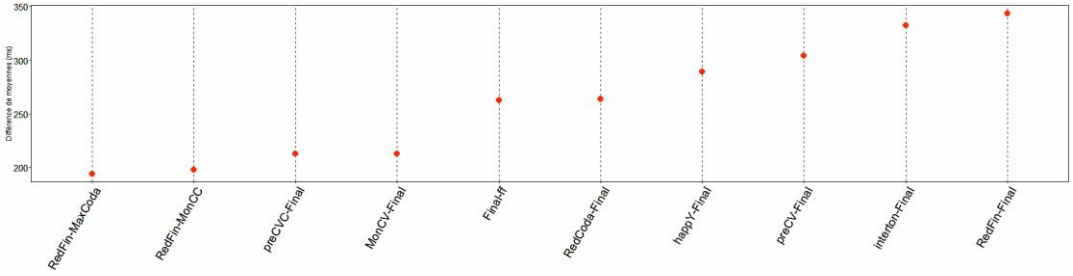


FIGURE 2 : Différences moyennes (ms) des 10 types de syllabes significativement différentes

Pour des questions de lisibilité, la figure 2 ne représente que les 10 premières valeurs significatives (en valeurs absolues) du test de comparaisons multiples sur les durées moyennes par type de syllabe. L'axe des abscisses représente les couples de syllabes significatifs par ordre croissant de pertinence. L'axe des ordonnées représente la différence moyenne entre type de syllabes (en millisecondes). Les différences les plus significatives entre les types de syllabes chez les anglophones opposent bien des syllabes inaccentuées à des syllabes accentuées ou finales. La durée des syllabes sous focus ne montre pas *a posteriori* la nécessité de bien distinguer cette catégorie dans notre annotation des types de syllabes et surtout de mettre cette propriété en tête dans nos critères d'annotation (voir section 3.2), elle est significativement différente de durée des monosyllabes en syllabe ouverte (monCV), mais pas des syllabes finales.

La position ne semble pas jouer entre les syllabes inaccentuées pré-toniques ou post-toniques. Les déformations privilégiées sont l'allongement des voyelles réduites et l'allongement plus important pour les syllabes ouvertes (syllabation en CV des langues romanes). La structure syllabique ne peut s'apprécier uniformément sur l'ensemble des types de syllabe considérés (la distinction n'est par exemple pas codée pour focus et Final) mais semble pertinente pour les syllabes accentuées.

Le statut de la position de la syllabe réduite dans le mot ne semble pas intervenir : la classification hiérarchique ascendante (fig. 3) des trois groupes en fonction des durées moyennes des types de syllabe montre l'indistinction pour les trois groupes de locuteurs entre preCV, ff et happY (syllabe inaccentuée avant accent, monosyllabe dans la chaîne parlée, ou syllabe après accent).

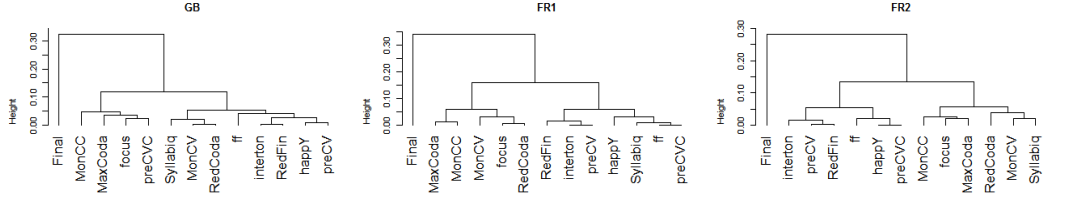


FIGURE 3 : classification ascendante des types de syllabe pour les natifs (GB), les FR1 et les FR2

Les voyelles réduites ont un comportement identique, qu'il s'agisse des finales ouvertes (happY ou des monosyllabes en forme faible). Les voyelles réduites intertoniques ne sont pas significativement différentes des voyelles prétoniques. Les syllabes les plus fortement accentuées (MaxCoda, focus) se rapprochent des monosyllabes non-réduites en syllabes fermées (MonCC). Les rares syllabes de fin d'énoncé (Final) sont les plus différenciées dans tous les groupes. Les syllabes qui font l'objet d'un focus sont réalisées comme des accentuées par les natifs et comme des syllabes réduites par les francophones.

5. Discussion et conclusion

5.1. Perspectives et prolongements

La méthode suivie ne suppose pas d'alignement préalable (mais n'a été testée que sur des enregistrements en chambre sourde). Elle n'est pas complètement automatique. Passée l'innovation technique, rien ne la distingue d'une annotation manuelle de l'ensemble du corpus qui extrairait par scripts l'ensemble des données. La différence notable avec Herment et al. (2014) est que la comparaison se fait effectivement sur les mêmes séquences de phones, alors que la syllabation peut varier.

Avantage de la méthode : l'aligneur s'affranchit de la structure de la syllabe phonétique (Meynadier 2001; Ridouane et al. 2011). Si l'on est ainsi privé de l'impression résultante (*comfortable* est réalisé en quatre syllabes par trois locuteurs francophones), la comparaison n'en reste pas moins fondée sur un patron comparable : l'intervalle segmental de la syllabe du modèle. La méthode ne permet pas de prédire les syllabations retenues par les locuteurs francophones (*comfortable* en 3 ou 4 syllabes). L'alignement sur la portion du spectre ne prédit pas la syllabe phonétique dans laquelle elle se trouve. Toutefois, le repérage des durées signale potentiellement des resyllabifications (transferts gabaritiques au sens de Ballier & Martin 2015).

Une analyse par classifieurs tels que TiMBL pourrait chercher à identifier les types de syllabes, voire les syllabes permettant le mieux de discriminer entre apprenants et natifs. La méthode de classification obtenue sur la base des métriques d'intervalles intervocaliques et interconsonantiques restant en dessous de 70% d'identifications correctes (Tortel 2009).

5.2 Conclusion

Nous avons privilégié la partie du corpus où les polysyllabes sont les plus nombreux, la différence entre les passages lus s'apprécie en fonction de la structure morpho-phonologique des mots : dans d'autres parties du corpus, les monosyllabes à syllabe fermée (monCVC) et ouverte (monCV) seraient surreprésentés.

Cette approche du discours des non-natifs est guidée malgré tout par un modèle, en l'occurrence, une certaine conception du découpage syllabique en anglais, qui n'est pas consensuelle. Il aurait été possible de procéder à un découpage ambisyllabique (Kahn 1976) de *terrified*, même si les dictionnaires du monde de l'édition ne suivent pas cette possibilité. Reste que l'importance accordée aux constituants prosodiques permet d'affiner la recherche de traits critériés (Hawkins & Filipović 2012) pour l'établissement de niveaux d'apprenants et de strates d'interlangue.

On pourrait étendre la comparaison des fragments de spectrogrammes à des unités prosodiques. En cas de consignes bien spécifiques, la méthode pourrait être étendue à des configurations prosodiques (ainsi, il se trouve que '*for one thing*' est réalisé avec un contour descendant, là où les francophones produisent plutôt une continuative mineure ascendante, sans que pour autant une montée soit *a priori* exclue chez les natifs). Typiquement, la réalisation du contour prosodique de *on* n'est pas identique entre la particule adverbiale et la préposition dans '*I didn't bother to switch the light on'* vs. *'tripped on a loose step in the dark*'. WinPitch LTL permet l'analyse en corpus du maintien ou non de la distinctivité des réalisations. Au sein de la variation interlocuteur, dans une perspective d'interlangue,

on cherche à dissocier le bon grain de la grammaire de l'ivraie de la variation phonostylistique. Cette approche centrée sur les durées pourrait être étendue aux événements prosodiques. Deux contours apparemment constants chez les anglophones pourraient être retenus : le *do* emphatique et l'incise parenthétique *worse still* en contour montant-descendant.

Remerciements

Nous remercions vivement Anne Tortel pour l'accès aux données de son corpus de thèse ANGLISH (Tortel 2009) ainsi que, pour leurs commentaires, les participants du workshop du 30 mars (Paris Diderot) ainsi que les relecteurs anonymes des JEP.

.

Références

ADAMS C. (1979). *English Speech Rhythm and the Foreign Learner*. Berlin: de Gruyter.

ARVANITI, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, 40(3), 351-373.

BALLIER, N. DELAIS-ROUSSARIE, E. HERMENT S. TORTEL A. (2014). Modélisation de l'intonation interlangue : le cas des questions. Actes des *JEP2014*

BALLIER, N. & MARTIN, Ph. (2015). Speech annotation of learner corpora. in GRANGER, S., MEUNIER, F. & GILQUIN, G. (eds) 2015 *Handbook of Learner Corpus Research*, Cambridge: CUP, 107-134.

BALLIER, N. (2016). La modélisation statistique du rythme et la dissolution de la structure syllabique, à paraître in BLANCKAERT, Cl., LÉON, J. & SAMAIN, D. (eds.), *Modèles et modélisations en sciences du langage, de l'homme et de la société. Perspectives historiques et épistémologiques*, Paris: L'Harmattan, 10 pages.

BELLMAN, R. (1957). *Dynamic Programming*. Princeton : Princeton University Press.

BIGI, B. (2012). SPPAS: a tool for the phonetic segmentation of speech. In *LREC* Vol. 8, 1748-1754.

CRUTTENDEN, A. (2015). *Gimson's Pronunciation of English*, Londres : Routledge.

CRYSTAL, D. (1996). The past, present and future of English rhythm. *Speak Out!* 18: 8-13.

DAUER, R. (1983). Stress-timing and syllable-timing reanalysed. *Journal of Phonetics* 11, 51–62.

DÍAZ-NEGRILLO, A., BALLIER, N., THOMPSON, P. (eds.). (2013). *Automatic treatment and analysis of learner corpus data*. (Studies in Corpus Linguistics 59). Amsterdam: John Benjamins Publishing Co.

GUT, U. (2003). Non-native speech rhythm in German. Proceedings of 15th International Congress of Phonetic Sciences, Barcelona. 2437–2440.

HAWKINS , J., FILIPOVIC, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework* (Vol. 1). Cambridge : Cambridge University Press.

HERMENT, S., DELAIS-ROUSSARIE, E., HERMENT S., TORTEL, A. (2014). Modelling interlanguage intonation: the case of questions, *Proceedings of the 7th Intenational Conference on Speech Prosody*, 20-23 mai 2014, Dublin, Irlande : 492-496

HUART, R. (2013) *Grammaire de l'anglais oral*, Paris : Ophrys.

KAHN, D. (1976). *Syllable-based generalizations in English phonology*. Ph.D. MIT.

Klatt, D. (1987). "Review of Text-to-Speech Conversion for English." *The Journal of the Acoustical Society of America* 82 (3), 737–93.

LOUKINA, A. KOCHANSKI, G. ROSNER, B. (2010). Rhythm measures and dimensions of durational variation in speech). *The Journal of the Acoustical Society of America*, 2011, vol. 129, no 5, 3258-3270.

MENNEN, I. (2007). Phonological and phonetic influences in non-native intonation, in TROUVAIN, J. & GUT, U. (eds.), *Non-native Prosody: Phonetic Descriptions and Teaching Practice*, 53–76, Mouton De Gruyter.

MENNEN, I. (2015). Beyond segments: towards an L2 intonation learning theory (LILT), in DELAIS-ROUSSARIE, E., AVANZI, M. & HERMENT, S. (eds.), *Prosody and languages in contact: L2 acquisition, attrition, languages in multilingual situations*, Springer Verlag, 176-188.

MEYNADIER, Y. (2001). La syllabe phonétique et phonologique: une introduction. *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA)*, *20*, 91-148.

RAMUS, F., NESPOR, M., MEHLER, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition* 73, 265–292.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

RIDOUANE, R., MEYNADIER, Y., FOUGERON, C. (2011). La syllabe: objet théorique et réalité physique. *Faits de langue 37*, 225-246.

ROACH, P. (1983). *English Phonetics and Phonology*, Cambridge: CUP.

TORTEL, A. (2009). *Evaluation qualitative de la prosodie d'apprenants français: apport de paramétrisations prosodiques*. Thèse de doctorat non publiée. Aix-Marseille University.

TORTEL, A. HIRST D. (2010). Rhythm metrics and the production of English L1/L2, Proceedings of Speech Prosody,

WELLS, J. (1990). *Longman Pronouncing Dictionary*, Londres: Longman.

WINPITCH (2016). WinPitch, www.winpitch.com

XU, Y., XU, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics*, 33:159-197.

XU, Y. (2013). ProsodyPro, A Tool for Large-scale Systematic Prosody Analysis, *Proceedings of the TRASP conference*, Aix-en-Provence.

Variabilité du geste palatal : effet du locuteur, de la structure syllabique et de l'accent sur différents types de consonnes en 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

RESUME

Les linguistes se sont régulièrement penchés sur la description du trait consonantique [+palatal] ; pourtant, le manque de données expérimentales constitue un obstacle au classement des consonnes concernées. Peu de travaux ont abordé la question du contrôle du geste lingual dans l'articulation palatale. Cependant, ils montrent que celui-ci semble bien plus complexe que dans d'autres consonnes. En russe, la plupart des consonnes possèdent une contrepartie palatalisée ce qui permet d'étudier les différences de réalisation du trait palatal au sein du même système. Nous proposons ici, à partir de données acquises avec un articulographe électromagnétique, de caractériser la variabilité du geste palatal impliqué dans la réalisation de différents types de consonnes palatalisées et pré-palatales du russe en fonction des facteurs locuteur, accent et structure syllabique.

ABSTRACT

Palatal gesture variability: speaker, stress and syllabic structure effects in Russian

Linguists have been studying palatal feature for a long time. Nevertheless, the lack of experimental data in the literature is a barrier to classification of consonants involving this feature. Few studies examined the question of the tongue gesture control in the production of [+palatal] consonants. However, they show that the tongue control process in palatal articulation seems more complex than in other types of consonants. In Russian, most of the consonants have a palatalized counterpart, in this way Russian language can be a reliable source for investigating different aspects of the palatal feature and its phonetic realization(s). Based on data acquired with an electromagnetic articulograph, we propose to characterize the palatal gesture involved in the production of different types of prepalatal and palatalized consonants in Russian and to observe the variability of this gesture by taking into account various factors such as speaker, stress and syllabic structure.

MOTS-CLES : Pré-palatal *vs* palatalisé, russe, variabilité, geste articulatoire, EMA

KEYWORDS: Prepalatal *vs* palatalized, russian, variability, aticulatory gesture, EMA

1 Introduction

La palatalisation est un processus de changement phonétique qui a pour particularité le déplacement du lieu d'articulation primaire vers la région du palais dur ou l'acquisition par une consonne d'une articulation secondaire palatale (Straka, 1965). Ce processus est à l'origine de plusieurs types de consonnes : (1) palatale, caractérisée par un geste d'élévation du dos de la langue vers le palais alors que la pointe de la langue est en position basse derrière les incisives inférieures – comme pour [c] produite par un geste articulatoire primaire d'occlusion palatale (Straka, 1965) ; (2) palatalisée, qui se caractérise par deux gestes et correspond à une articulation élaborée (Lindblom & Maddieson, 1988), c'est-à-dire qu'à la constriction ou obstruction principale se superpose une deuxième

constriction. L'articulation secondaire se traduit donc par la présence d'un deuxième rétrécissement dans le conduit vocal d'aperture plus large que la constriction primaire. La palatalisation relèverait donc d'une articulation secondaire. Ladefoged & Maddieson (1996) évoquent la présence d'un timing court et précis entre les deux articulations.

La description du geste palatal reste un sujet de débat (Keating, 1988 ; Recasens, 1990) même si des études l'ont abordée sous différents aspects (Kuznetsova, 1969 ; Kochetov, 2002 ; Kedrova et al., 2008) alors que les caractéristiques phonétiques sont beaucoup plus stables pour les consonnes labiales, vélaires, dentales. Peu de travaux se sont penchés sur la question du contrôle du geste lingual dans la production des consonnes [+palatal]. Ils montrent cependant que le processus de contrôle de la langue y semble bien plus complexe que dans d'autres consonnes comme dentoalvéolaires ou vélaires (Lindblom & Maddieson, 1988 ; Recasens & Romero, 1997). De plus, les résultats existants divergent quant à la caractérisation du geste lingual entre consonnes palatalisées et palatales. D'après Recasens & Romero (1997), les palatales sont des articulations simples (au contraire des palatalisées), donc sans superposition de deux articulations (principale et secondaire). Inversement, Keating (1988) les considère comme des articulations complexes impliquant dans leur production deux gestes articulatoires (la lame et le dos de la langue) et donc, deux contrôles moteurs distincts. L'examen des divergences dans la description de la nature articulatoire des consonnes [+palatal] montre que peu d'études antérieures se sont penchées sur la variabilité des réalisations, hormis la variabilité liée aux différences de contexte vocalique (ex. Recasens et al., 1993).

Le système phonologique du russe, comme celui d'autres langues slaves, comporte de nombreuses consonnes palatalisées alors que les consonnes palatales et palatalisées, quand elles sont phonologiques, sont peu fréquentes dans les langues du monde (selon Maddieson (1984), en moyenne, au plus 10 % des inventaires consonantiques) et peuvent être extrêmement rares au regard de certains modes, comme par exemple les plosives, latérales et nasales *vs* glides (Vallée et al., 2002). Le russe est particulièrement intéressant du fait de cette spécificité : la plupart des consonnes possède une contrepartie palatalisée, communément appelée consonne molle par opposition aux consonnes dites dures. En russe, il est aisé de trouver des paires minimales avec consonne palatalisée : /val/~/vʲal/ *digue~mou (forme courte)*, /brat/~/bratʲ/ *frère~prendre*, /luk/~/lʲuk/ *oignon~trappe*. D'autre part, la palatalisation peut être allophonique due à une assimilation régressive d'une consonne palatalisée sur la consonne qui précède : /snʲeg/ > [sʲnʲek] *neige*, /zdʲesʲ/ > [zʲdʲesʲ] *ici*. Enfin, le système possède également les consonnes pré-palatales /ʧ ʃ ʃː ʒ ʒː/. Le choix du russe est donc pertinent pour étudier les différences de réalisation du trait *palatal* au sein du même système.

L'étude pilote que nous proposons vise à caractériser le geste palatal impliqué dans la réalisation des différents types de consonnes [+palatal] du russe et d'observer la variabilité de ce geste en prenant en compte différents facteurs : locuteur, accent lexical, structure syllabique. Il s'est agi d'examiner, selon ces facteurs, la forme du geste lingual lors de l'atteinte de la cible consonantique ainsi que le timing du geste lingual, c'est-à-dire l'organisation temporelle de quatre points de référence sur la langue pour atteindre cette cible.

Selon Browman & Golstein (1995) ou Byrd (1995), les différentes positions intra-syllabe correspondent à différents modes de configuration des gestes : en attaque de syllabe les gestes consonantiques sont coordonnés en phase, alors qu'en coda est observée une moins grande stabilité dans la coordination des gestes du noyau vocalique et de la consonne, ce qui expliquerait, selon Browman & Golstein, la vulnérabilité de la coda sujette à la lénition dans de nombreuses langues (Kingston, 2008). Notre objectif est de caractériser les différences dans le timing du geste palatal en russe en fonction de la position dans la syllabe. Comme l'accent est souvent cité dans la littérature comme une cause de renforcement articulatoire (ex. Fougeron & Keating, 1997 ; Krakow, 1999), les différences dans le timing seront de la même manière observées sous l'accent et en position atone.

2 Méthodologie

2.1 Hypothèses

L'examen des travaux antérieurs permet de poser les hypothèses suivantes : (1) le timing reflète le type primaire ou secondaire de l'articulation : si l'articulation palatale est secondaire, l'élévation de la partie antérieure de la langue 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) la palatalisation sous l'accent subit un renforcement articulatoire et tend vers une articulation palatale (Straka, 1963 ; Fougeron, 1998) ; (3) le timing des gestes est plus synchrone en position d'attaque qu'en position de coda (Browman & Goldstein, 1988, 1995 ; Byrd, 1995).

2.2 Participants et stimuli

Trois locutrices âgées de 27, 28 et 42 ans de langue maternelle russe, originaires respectivement de Sysert (Oural), Mourom (région de Vladimir) et Dzerzhinsk (région de Nizhnij Novgorod), ont participé à cette expérience pilote.

Un ensemble de 40 mots contenant une des consonnes cibles forme le corpus et inclut entre autres des paires minimales comportant une consonne [-palatal] *vs* une consonne palatalisée. Les consonnes [+palatal] apparaissent en position accentuée *vs* atone, et attaque *vs* coda. Dans cet article sont présentés les résultats préliminaires pour les 5 consonnes /t tʲ ʧ ʃ ʃː/ (table 1).

Paire minimale	Accentuée *vs* atone		Attaque *vs* coda
	Paire minimale		
/tuk/ ~ /tʲuk/ *bruit provoqué par un coup sur une surface dure* vs *gros sac*	/ʃːit/ *vs* /ʃːi.ˈta/ *bouclier* vs *bouclier (génitif, sg.)*	/ˈtʲe.la/ ~ /tʲeˈla/ *corps (génitif, sg.)* vs *corps (pluriel)*	/tʲap/ *vs* /matʲ/ *faire qch à la va-vite* vs *mère*
/ʃestʲ/ *vs* /ʃesˈtʲi/ *six* vs *six (génitif, datif, locatif)*	/ˈʧer.tʲi/ ~ /ʧerˈtʲi/ *diablotin (pluriel)* vs *tracer (injonction)*	/ʧem/ *vs* /meʧ/ *conjonction de subordination* vs *épée*	
		/ʃːeli/ *vs* /lʲeʃːi/ *fente* vs *brème*	
		/ʃov/ *vs* /voʃ/ *suture* vs *poux*	

TABLE 1 : Oppositions phonologiques et contrastes observés.

Chacun des mots comportant une des consonnes cibles a été inclus dans une phrase porteuse /tɨ ˈvʲi.dʲe.la/ *cible* /dva (trʲi) ˈra.za/ 'tu as vu '*cible*' deux (trois) fois' facilitant le repérage des frontières lors de la segmentation. Les énoncés ont été présentés 5 fois dans un ordre aléatoire pour chaque locutrice. La consigne a été donnée de lire à voix haute les énoncés à un débit de parole normal.

2.3 Protocole

Les enregistrements se sont déroulés dans la chambre anéchoïde du Gipsa-lab. Les données ont été enregistrées avec le système AG 200 de la société Carstens (2D) à une fréquence d'échantillonnage de 200 Hz. Quatre bobines ont été collées sur la langue : apex, lame, sommet du dos de la langue lorsque celle-ci est élevée et arrondie et une un peu plus en arrière de cette position ; une bobine collée sur les incisives inférieures afin de récupérer le mouvement mandibulaire. Deux bobines de référence (nez, incisives supérieures) ont été rajoutées afin de corriger les mouvements de la tête dans le plan médio-sagittal des sujets. Les données acoustiques ont été recueillies avec un micro AKG C1000S et un enregistreur numérique PMD 670 (échantillonnage à 22,05 kHz). Les locutrices équipées de l'articulographe étaient assises face à un écran où s'affichaient les énoncés à lire.

2.4 Mesure et analyses

Les trajectoires des articulateurs ont été analysées avec un logiciel interne (TRAP) développé sous Matlab au Gipsa-lab. Sur les trajectoires des déplacements des différentes bobines ont été repérés automatiquement les minima et maxima à partir des passages par zéro de la courbe de vitesse de chacun des articulateurs. Les cibles articulatoires /t tʲ ʧ ʃ ʃː/ ont été considérées atteintes lorsque l'apex était à sa position maximale. La forme linguale à l'instant de l'atteinte de la cible

consonantique était donnée par les positions en X (degré d'avancement dans la cavité buccale) et Y (hauteur de la langue) des trois points mesurés : lame, mid- et post-dorsum, lorsque l'apex avait atteint sa position maximale. Les données articulatoires de chaque locutrice ont été recalées dans leur plan occlusal X/Y respectif ayant pour origine la bobine la plus antérieure de celui-ci.

Pour analyser le timing du geste lingual, nous avons observé la coordination temporelle des quatre bobines dans la production des cinq consonnes. Pour chacun des quatre points de la langue, l'écart temporel entre l'instant de référence et l'instant où chaque point de la langue atteint son maximum a été mesuré. Nous avons choisi comme instant 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. Ce point est l'instant à partir duquel la mandibule remonte pour réaliser l'articulation consonantique.

Nous souhaitons étudier les variations d'une variable dépendante *durée* (ms) et l'influence de plusieurs facteurs sur celle-ci déclinés en plusieurs modalités : (1) consonne /t tʲ ʧ ʃ ʃː/, (2) syllabe tonique *vs* atone, (3) attaque syllabique *vs* coda. Compte tenu de ces éléments et du fait que pour chaque mot cible plusieurs valeurs de durée ont été mesurées pour un même sujet, nous avons effectué les analyses statistiques en utilisant un modèle linéaire à effet mixte à l'aide de la fonction lme du package nlme du logiciel R (Bazzoli et al., 2015). Pour étudier un éventuel impact de la force articulatoire[1], nous avons choisi de créer une variable *renforcement articulatoire* avec trois modalités : (1) attaque de syllabe atone (Atone∗Onset), (2) attaque de syllabe accentuée (Accent∗Onset), (3) coda de syllabe accentuée (Accent∗Coda).

Deux analyses distinctes ont été réalisées : l'une avec les seuls /t/ et /tʲ/ dans la modalité Accent∗Onset, l'autre sans la consonne /t/. Les mesures prises simultanément sur les quatre points de la langue font qu'elles sont fortement corrélées les unes avec les autres. Pour en tenir compte, nous avons utilisé les paramètres Weights et Correlation de la fonction lme. À partir du modèle ont été effectuées des analyses de comparaisons multiples de moyennes après avoir construit les matrices de contrastes adéquates en utilisant la méthode de Hothorn et al. (2008) et en utilisant la fonction glht du package multcomp de R. L'objectif des comparaisons multiples était de décrire le timing des quatre bobines lors de la production d'une consonne donnée et de comparer les différences entre consonnes en fonction des trois modalités du facteur *renforcement articulatoire*.

3 Résultats

L'objectif de la présente étude est de caractériser la variabilité du geste palatal impliqué dans la réalisation des consonnes /t tʲ ʧ ʃ ʃː/ du russe. Dans les figures suivantes sont représentées, par locutrice, les positions des quatre bobines pour les cinq répétitions du mot cible.

3.1 Contours linguaux

/tuk/~/tʲuk/

La figure 2 montre les résultats obtenus sur la mesure des points lame, mid- et post-dorsum au moment où l'apex est dans une position maximale pour /t/ et /tʲ/ dans la paire minimale /tuk/~/tʲuk/. Les locutrices IH et IM produisent /t/ dans /tuk/ en élevant l'apex avec une configuration en X des autres points quasi alignée montrant qu'il s'agit d'une consonne apicale produite dans la région alvéolaire. Le mode articulatoire de /t/ est aussi apical chez la locutrice KB avec toutefois les positions du mid- et du post-dorsum plus élevées que celles des bobines de la partie antérieure de la langue. Des différences entre IH, IM et KB sont trouvées dans /tʲuk/. La production de la consonne palatalisée /tʲ/ par IH montre une élévation de la lame ainsi qu'une élévation du mid-dorsum. Comme pour /t/, la locutrice IM produit /tʲ/ avec la même forme de la langue que IH ; à la différence de KB qui utilise une autre stratégie impliquant une élévation (+ 6 mm) du mid-dorsum.

[1] Selon Straka (1963 : 61), « l'énergie du mouvement articulatoire c'est la force avec laquelle les muscles de la langue et du maxillaire se contractent en vue de mettre en place ces organes pour l'articulation voulue. »

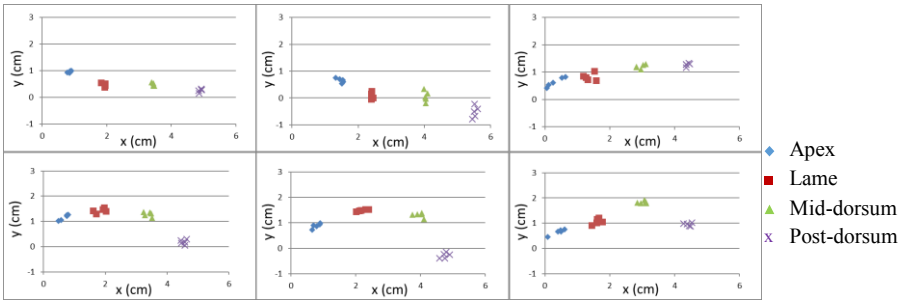


FIGURE 2 : Localisation des 4 bobines dans la cavité buccale des 3 locutrices IH (à gauche), IM (au milieu) et KB (à droite) pour la paire minimale /tuk/ (en haut) ~ /tʲuk/ (en bas) dans le plan X/Y du sujet.

/ˈtʲe.la/~/tʲe.ˈla/

La figure 3 montre que IH et IM présentent des patrons proches pour les consonnes palatalisées de cette paire minimale avec toutefois plus de variabilité au niveau du mid- et du post-dorsum chez IM : /tʲ/ est apical chez IM et plutôt apico-laminal chez IH. Chez ces deux locutrices, on n'observe pas de différences dans la production de /tʲ/ en initiale de syllabe tonique et atone alors qu'on aurait pu s'attendre à ce que le dos de la langue soit plus élevé dans la syllabe tonique (Straka, 1963 ; Fougeron & Keating, 1997). Chez KB l'apex est abaissé, le mid-dorsum élevé. On observe moins de différences dans l'élévation de l'apex, de la lame et du mid-dorsum dans /tʲe.ˈla/ ce qui donne à la langue une forme globale plus bombée. Ce résultat pour les trois locutrices montre que l'accent a visiblement peu d'effet sur la forme de la langue lors de l'atteinte de la cible articulatoire : les quatre points localisés sur la langue sont similaires en positions tonique et atone.

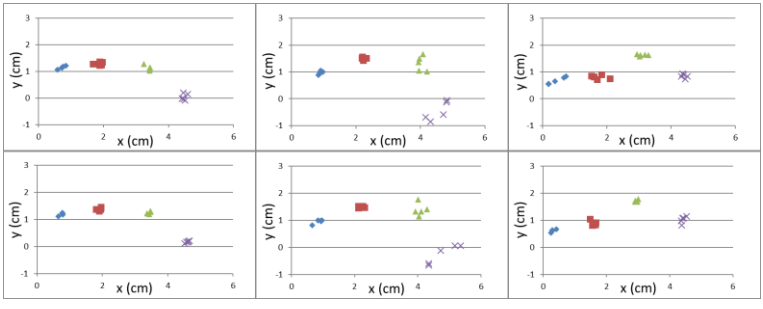


FIGURE 3 : Localisation des 4 bobines dans la cavité buccale des 3 locutrices IH (à gauche), IM (au milieu) et KB (à droite) dans /ˈtʲe.la/ (en haut) ~ /tʲe.ˈla/ (en bas).

Les mêmes données ont été obtenues pour /ʧ ʃ ʃʲ:/ et les mêmes observations sont faites en syllabes tonique et atone. Pour ces consonnes aucune différence de la forme linguale liée à l'accent n'a été observée à l'atteinte de la cible articulatoire. Pour /ʧ/, on observe plus de variabilité entre les cinq répétitions en X (≈ 3 mm) sur chaque position dans la configuration linguale lors de la production de la syllabe atone chez IH et IM et tonique (≈ 4 mm) chez KB.

/tʲap/ *vs* /matʲ/

Figure 4 sont présentées les positions des quatre points de la langue pour /tʲ/ en attaque *vs* coda dans /tʲap/ *vs* /matʲ/. Pour IH /tʲ/ est apico-laminale. On observe moins de variabilité au niveau du lieu d'articulation entre les cinq répétitions en coda et plus de variabilité (≈ 3,5 mm) en attaque. L'articulation est plutôt apicale dans les deux cas chez IM avec un contour global bombé de la langue lorsque la consonne cible est en coda. Chez KB l'articulation est dorsale. Le post-dorsum est plus en arrière et plus bas en coda. Les faibles différences relevées entre les positions attaque et coda peuvent s'expliquer par le fait que la consonne cible se trouve dans des mots monosyllabiques et donc sous l'accent.

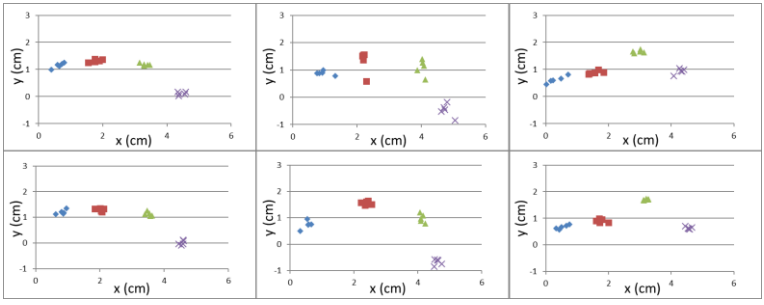


FIGURE 4 : Localisation des 4 bobines dans la cavité buccale des 3 locutrices IH (à gauche), IM (au milieu) et KB (à droite) pour /tʲap/ (en haut) *vs* /matʲ/ (en bas).

Pour /ʧ/, on relève de la variabilité entre les 5 répétitions en X lorsqu'elle est initiale ou finale de syllabe chez IH et IM et plus de variabilité chez KB en attaque. La configuration de /ʃ/ est stable entre les différentes répétitions chez KB pour l'attaque et la coda alors que chez IH et IM est observée plus de variabilité en X (≈ 2,5 mm pour IH et ≈ 4 mm pour IM) et en Y (≈ 4 mm pour ces deux locutrices) lorsque la consonne est en coda. Nous avons observé plus de variabilité entre les répétitions de /ʃʲ:/ réalisées en attaque de syllabe chez les locutrices IM et KB et de la variabilité en X chez IH que ce soit attaque ou coda de syllabe. On n'observe pas de différences dans la production de /ʧ ʃ ʃʲ:/ entre les positions attaque et coda alors qu'on aurait pu s'attendre à la réduction de l'amplitude du mouvement articulatoire en coda.

3.2 Timing du geste lingual

L'objectif est de décrire et comparer l'organisation temporelle des quatre localisations sur la langue lors de la production des consonnes /t tʲ ʧ ʃ ʃʲ:/ en fonction des trois modalités du facteur *renforcement articulatoire* qui combine les facteurs accent et position dans la syllabe (section 2.4).

Modalite Accent*Onset

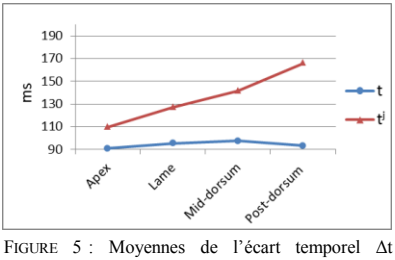


FIGURE 5 : Moyennes de l'écart temporel Δt estimées par la méthode des moindres carrés pour les 4 points de localisation sur la langue pour les consonnes /t/ et /tʲ/ dans la modalité Accent*Onset.

Pour /t/, il n'y a pas de différences significatives trouvées entre les Δt moyens mesurés pour chacun des points sur la langue. Rappelons que Δt correspond à l'écart temporel entre le minimum atteint par la mandibule dans la réalisation de la voyelle qui précède et le maximum atteint par chaque point de la langue. Cela suggère un contrôle simultané des quatre zones de la langue et donc un contrôle global du geste lingual (figure 5). Pour /tʲ/, l'écart temporel est significatif entre le geste apical (geste de base) et le geste palatal (respectivement pour mid-dorsum et post-dorsum : z=5,84 p<0,01 ; z=10,02 p<0,01) : l'apex atteint le lieu d'articulation situé dans la zone dentoalvéolaire en premier suivi du mid-dorsum et du post-dorsum. La comparaison entre /t/ et /tʲ/ indique que les Δt pour mid-dorsum et post-dorsum ainsi que pour la lame de la langue sont significativement différents (z= - 4,46 p<0,01 ; z= - 7,24 p<0,01 ; z= - 3,36 p=0,01). La palatalisation affecte donc le timing de l'ensemble du geste lingual et l'apex anticipe de 56 ms en moyenne (σ=35ms) l'élévation du post-dorsum.

Nous avons ensuite observé les variations de la variable Δt en fonction du facteur renforcement articulatoire et du facteur consonne /tʲ ʧ ʃ ʃʲ:/. Dans la modalité Accent*Onset pour /ʧ/ (figure 6), le patron du timing lingual est similaire à celui de /tʲ/. On constate d'abord une élévation de la pointe suivie d'une élévation du dos de la langue (mid-dorsum et post-dorsum, respectivement z=7,21 p<0,01 ; z=10,69 p<0,01) ce qui confirme l'articulation apicale et apico-laminale de ces consonnes chez IH et IM relevée dans l'étude sur la forme linguale au moment de l'atteinte de la cible. Toutefois, le geste dorsal est plus lent pour /ʧ/ que pour /tʲ/. L'explication se trouve peut-être du côté

de la durée acoustique car l'affriquée [ʧ] est composée d'une partie fricative [ʃ]. La réalisation d'une fricative demande un contrôle plus fin de la constriction pour générer le bruit de friction (Vallée et al., 2002). Pour /ʃ/ (figure 6), la pointe et la lame s'activent d'abord suivies de l'élévation du post-dorsum. Pour /ʃː/, l'élévation du mid-dorsum est suivie de l'élévation du post-dorsum. Le geste articulatoire est en moyenne significativement plus lent pour /ʃ/ et /ʃː/ contrairement à /tʲ/ ($p<0,05$).

Modalité Atone∗Onset

Les consonnes /tʲ/ et /ʧ/ qui possèdent le trait [+palatal] montrent que l'activation de la pointe de la langue précède celle du mid-dorsum pour /tʲ/ et celle du post-dorsum pour /ʧ/ (figure 6). En revanche pour /ʃ/ les quatre points s'activent ensemble. Pour /ʃː/, on constate d'abord une élévation du mid-dorsum suivie d'une élévation simultanée de la lame et du post-dorsum alors que l'étude de la forme linguale montre que chez les locutrices IH et IM la constriction est apico-laminale et qu'elle est dorsale chez KB. Des différences sont donc observées par rapport à la modalité Accent∗Onset : on relève ainsi un timing plus synchrone des quatre points, et donc une activation moins différenciée, entre les quatre parties de la langue pour /tʲ ʧ ʃ/ dans la modalité Atone∗Onset. Au contraire, pour /ʃː/, on note une activation moins différenciée en syllabe accentuée. Les différences significatives dans le timing pour une même localisation entre les différents types de consonnes coronales sont moins nombreuses dans cette modalité par rapport à Accent∗Onset. On peut noter qu'entre /tʲ/ et /ʃː/, les Δt concernant la pointe et la lame de la langue sont plus courts pour /tʲ/ ce qui indique que la phase d'atteinte de la cible /ʃː/ est plus longue que la phase d'atteinte de /tʲ/. Cette différence a également été observée sous l'accent et reste cohérente avec les propositions de Vallée et al. (2002). Pour ce qui concerne /ʃ/ et /ʃː/, la lame atteint son maximum plus rapidement pour /ʃ/ ($z= - 3,84$ $p=0,02$). Ce résultat rejoint la forme de la langue observée lors de l'atteinte de la cible articulatoire qui montre une position plus basse de la lame pour /ʃ/ que pour /ʃː/ chez les deux locutrices IM et IH. C'est le contraire pour KB (lame plus basse pour /ʃː/ que pour /ʃ/).

Modalité Accent∗Coda

Pour cette modalité (figure 6), on n'observe pas d'écart significatif dans le timing des quatre localisations lors de la production d'une consonne donnée ce qui suggère un timing plus synchrone des quatre points et une activation moins différenciée entre les parties de la langue lorsque la consonne est en coda. De plus, on ne relève pas de différences significatives dans le timing d'un même point entre les différents types de consonnes. L'effet de la position en coda affecte donc plus la dynamique du geste que la forme de la langue.

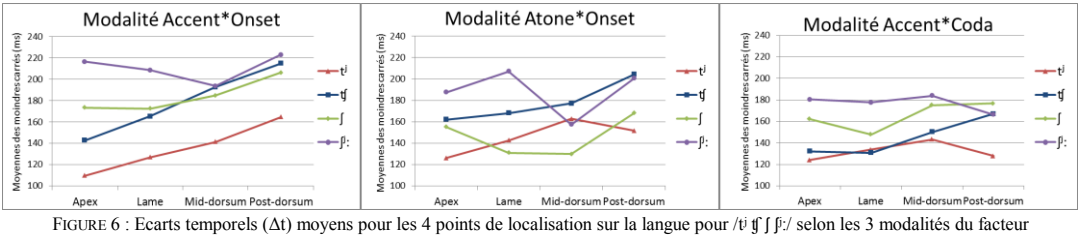


FIGURE 6 : Ecarts temporels (Δt) moyens pour les 4 points de localisation sur la langue pour /tʲ ʧ ʃ ʃː/ selon les 3 modalités du facteur *renforcement articulatoire.*

4 Discussion

Les résultats de cette étude pilote suggèrent que la variabilité dans les productions relèverait davantage du locuteur que de la position dans la syllabe ou de l'accent. L'analyse des quatre points montre deux formes articulatoires pour la cible /t/. La première consiste à relever l'apex, avec lame, mid- et post-dorsum en position basse. Dans la deuxième réalisation, la pointe de la langue est abaissée. Des différences sont relevées dans la palatalisation de /t/. La première des réalisations consiste à abaisser l'apex et à élever la lame et dans une moindre mesure le dos, la masse de la langue est alors plus antérieure pour /tʲ/. Pour la seconde réalisation, c'est l'élévation du dos de la langue qui est observée avec une forme bombée permettant de rapprocher l'articulation de /tʲ/ de

celle de la consonne [c] décrite par Straka (1965) qui affirmait que pour atteindre la zone la plus haute, les muscles élévateurs sont sollicités pour permettre une poussée verticale importante de la langue. L'élévation de la langue a des répercussions sur l'apex : celui-ci se dirige alors vers les incisives inférieures. Cet élément constitue une des particularités des consonnes palatales d'après Straka : pour [t], [d] ou [n] produites habituellement avec l'apex relevé, le fait qu'il soit abaissé est capital pour la palatalisation. Cependant, KB produit la cible /t/ apicale avec l'apex abaissé. Lors de la palatalisation, la pointe de la langue étant déjà basse, il suffit au dos de la langue de s'élever vers le palais dur. Nos résultats suggèrent que si l'apex est abaissé dans [t], la palatalisation s'effectue avec une élévation du dos sans modifier la forme de la partie antérieure de la langue ; si [t] est produite avec la pointe relevée c'est la position de la lame qui apparait déterminante pour la palatalisation. Dans ce dernier cas le degré de palatalisation semble moins fort. On peut penser que l'apex relevé dans la production d'une articulation primaire freine le degré de la palatalisation. Ces deux stratégies observées pour la production de palatalisées rappellent les résultats de Kedrova et al. (2008) qui montrent un premier type de réalisation avec une activité de la lame en contact avec la région post-alvéolaire, ce qui correspond à la stratégie décrite comme laminale chez IH et IM, et une seconde réalisation qui implique une très forte élévation du dos de la langue et l'avancement de celle-ci dans la cavité buccale, ce qui se rapproche de ce que nous avons observé chez KB, cependant sans antériorisation de la masse de la langue.

Nos résultats sur la forme de la langue ne montrent pas de différences selon l'accent et la position dans la syllabe pour les trois locutrices. Notre hypothèse d'un geste lingual plus ample et plus stable dû au facteur renforcement articulatoire, n'est pas validée par nos observations.

Dans un second temps, nous avons analysé le timing des quatre bobines pour /t tʲ ʧ ʃ ʃʲ:/ en fonction des trois modalités observées. Pour la modalité Accent∗Onset, un décalage temporel est observé chez tous les locuteurs entre l'articulation apicale et l'articulation secondaire palatale qui arrive plus tard. Ces observations sont conformes à celles de Ladefoged & Maddieson (1996). Concernant les palatalisées et palatales /tʲ ʧ ʃ ʃʲ:/, des particularités phonétiques sont relevées. Dans la modalité Accent∗Onset, des différences significatives ont été trouvées (1) dans le timing d'un même point de localisation entre les différents types de consonnes (2) dans le timing des quatre localisations lors de la production d'une consonne donnée. Ces différences sont moins souvent significatives dans la modalité Atone∗Onset et jamais significatives dans la modalité Accent∗Coda. Connue pour être une position lénifiante, la coda a tendance à gommer les différences dans la dynamique du geste lingual pour l'ensemble des consonnes observées alors que pour ces mêmes consonnes, la forme de la langue lors de l'atteinte de la cible articulatoire ne montre pas de différences. Ce résultat gagnera en précision avec une extension de l'analyse intégrant la structure /ˈCV.VC/ et donc la modalité Atone∗Coda. Contrairement aux propositions de Browman & Goldstein (1995 : 26), nos données ne suggèrent pas une réduction de l'amplitude du mouvement articulatoire en coda qui serait causée par une baisse de l'effort articulatoire ou une coopération des différents points de la langue ayant pour but de produire un geste moins extrême en coda. De même, on ne retrouve pas de manière régulière chez un même locuteur, ni pour une même consonne, d'écart temporel plus court en attaque qu'en coda entre les maxima atteints par les quatre points de la langue qui pourrait suggérer une cohésion articulatoire plus forte en attaque de syllabe (Browman & Goldstein, 1988, 1992).

Les intervalles temporels mesurés ne sont pas significativement différents pour une même consonne et un même point de localisation sur la langue en fonction des modalités observées. L'accent ou la position dans la syllabe semble donc ne pas avoir d'effet sur le timing d'un point de la langue pour une même consonne. En revanche, la significativité des différences d'intervalles temporels relevées entre les points de localisation sur la langue pour une même consonne en syllabe accentuée *vs* syllabe atone n'a pas encore été testée. Ce prolongement de l'analyse permettra peut-être de vérifier une synchronisation articulatoire plus forte entre les différentes parties de la langue sous l'accent et donc un effet de l'accent sur la dynamique du geste plutôt que sur la forme linguale à l'atteinte de la cible.

Remerciements

Cette étude pilote est réalisée dans le cadre du Projet ANR-10-BLAN-1916 APPSy.

Références

BAZZOLI C., LETUE 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.

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].

HOTHORN T., BERTZ F. & P. WESTFALL (2008). Simultaneous inference in general parametric models. *Biometrical journal* 50(3), 346-363.

KEATING P. (1988). Palatals as complex segments: X-ray evidence. *UCLA Working Papers in Phonetics* 69, 77-91.

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.

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.

KRAKOW R. A. (1999). Physiological organization of syllables: a review. *Journal of Phonetics* 27, 23-54.

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. PALLARES (1993). An electropalatographic study of alveolar and palatal consonants in Catalan and Italian. *Language and Speech* 36(2-3), 213-234.

RECASENS D. & J. ROMERO (1997). An EMMA study of segmental complexity in alveolopalatals and palatalized alveolars. *Phonetica* 54, 43-58.

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., 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.

Variation prosodique et traduction poétique (LSF/français) : Que devient la prosodie lorsqu'elle change de canal ?

Fanny Catteau[1, 2], Marion Blondel[1, 2], Coralie Vincent[1, 2],
Patrice Guyot[2, 3], Dominique Boutet[2, 4]

(1) SFL, CNRS-Paris8 59 rue Pouchet, 75017 Paris, France
(2) Labex ARTS-H2H, Paris8, 2 rue de la Liberté, 93000 Saint Denis, France
(3) IRIT, Toulouse III, 118 Route de Narbonne, 31062 Toulouse, France
(4) LIAS/IMM (EHESS & CNRS), Université Evry-Val-d'Essonne
fannycatteau90@gmail.com, marion.blondel@cnrs.fr

RESUME

L'étude de la prosodie des langues vocales repose en partie sur la mesure des paramètres de durée, d'intensité et de fréquence sonores. Les langues des signes, quant à elles, empruntent le canal visuo-gestuel et mobilisent des articulateurs manuels et non manuels (buste, tête, éléments du visage). Notre étude a pour objectif d'établir des outils permettant de comparer, au niveau prosodique, la traduction en français de séquences poétiques et la version originale en langue des signes française (LSF). Nous avons recueilli des données vidéo augmentées de capture de mouvement – qui offrent plusieurs pistes d'exploration des paramètres prosodiques pour la LSF – ainsi que des données audio des traductions en français – qui révèlent les stratégies des interprètes pour interpréter la variation prosodique.

ABSTRACT

Prosodic variation and poetic translation (LSF/French): What happens to prosody with a channel change?

The study of spoken language prosody is partially based on the measurement of three parameters: sound duration, intensity and frequency. Sign languages use the visual-gestural channel and mobilize manual and non manual articulators (chest, head, facial features). Our study aims to establish tools that allow to compare, on the prosodic level, the translation of poetic sequences into French and their original versions in French Sign Language (LSF). We collected video and motion capture data – which offers wide avenues for exploration of LSF prosodic features – as well as audio data of French translations – that reveal the strategies of interpreters to convey prosodic variation.

MOTS-CLES : prosodie, langues des signes, modalité, traduction, capture de mouvement
KEYWORDS : prosody, sign languages, modality, translation, motion capture

1 Comment appréhender la prosodie gestuelle

L'étude du segmental ou du suprasegmental peut être menée tant dans la modalité visuo-gestuelle que dans la modalité audio-vocale. Ainsi, des paramètres tels que la durée, la fréquence ou l'intensité devraient être opératoires tant pour l'étude du flux sonore que pour celle du flux gestuel. Pour autant, la recherche en prosodie des langues des signes hésite encore sur ce qui constitue le niveau suprasegmental. La question du statut du non-manuel (buste, tête, mimique) est rarement évoquée, tout comme les comparaisons strictes entre les paramètres intonatifs des langues vocales (intensité, durée, fréquence) et leurs équivalents dans le canal gestuel.

Le registre poétique nous semble un espace privilégié pour étudier les phénomènes prosodiques, et la traduction du registre poétique, d'une modalité à l'autre, paraît être une piste expérimentale fructueuse. Il s'agit également d'une approche innovante : le patrimoine poétique en LSF étant de taille très modeste, et surtout peu diffusé, les occurrences de traduction vers le français sont plus rares que celles vers la LSF, et n'ont pas été étudiées, à notre connaissance, jusqu'aux travaux récents de Catteau (2015) et Corominas (2015).

A travers l'expérience du projet CIGALE[1], et en particulier de son extension sous la forme d'une expérimentation de traduction poétique, nous avons plusieurs objectifs : rechercher les points communs ou les spécificités de chacune des modalités quant à leur prosodie et établir des mesures basées sur les données biomécaniques. Au final, nous confirmons que l'intonation gestuelle s'appuie bien sur des modulations de l'intensité, de la durée et de la fréquence (restant à préciser en lien avec l'étude de la vitesse et de ses dérivées) et que nous pouvons distinguer des *motifs* ou contours prosodiques systématiques en nous appuyant sur les régularités et contrastes rythmiques et 'spatio-mélodiques', autrement dit une utilisation régulière et contrastée de l'espace (trajectoire répétées, zones mobilisées selon un axe de symétrie, emplacements opposés sur des axes précis).

Dans un premier temps, nous ferons un état de la recherche en prosodie dans les langues à modalité visuo-gestuelle. Ensuite, nous présenterons le cadre de l'expérimentation que nous avons mise en place, ainsi que le projet dans lequel elle s'inscrit. Enfin, dans un dernier temps, nous présenterons les résultats obtenus grâce à notre étude et les perspectives envisagées.

1.1 Prosodie audio-vocale et prosodie visuo-gestuelle

Les recherches sur la prosodie dans les langues des signes sont récentes et peu diffusées (relativement à la prosodie des langues vocales). Elles ne s'attellent que rarement à l'étude de ces trois paramètres (durée, fréquence, intensité). Il a été proposé à plusieurs reprises (par exemple Sandler, 1999) que l'équivalent de l'intonation des langues vocales était essentiellement non manuel (inclinaison de la tête, du buste) et en particulier lié aux articulateurs faciaux (mouvements des sourcils, mouvements des joues, de la bouche, entre autres). Pour d'autres auteurs, l'équivalent de l'intonation vocale ne peut se limiter au non-manuel et les articulateurs non manuels ne sont pas non plus spécifiques aux signeurs. Ainsi, Blondel & Le Gac (2007) ont souligné, pour la LSF, que les expressions faciales, et plus généralement le non-manuel, peuvent jouer un rôle essentiel dans la signification d'un énoncé en langue des signes, tout en étant également présents chez les entendants dans une communication en face à face.

[1] http://www.labex-arts-h2h.fr/cigale-104

1.1.1 *La place du non-manuel*

Parce que la dimension incarnée des langues vocales est étudiée de plus en plus finement (cf. Ferré, 2014, entre autres), certaines attributions des langues des signes peuvent être reversées dans le pot commun de la gestualité humaine et de la dimension multimodale du langage en interaction. Ainsi, la caractérisation 'linguistique' des mouvements faciaux en langue des signes reposait dans un premier temps sur l'ajustement précis de leur portée temporelle à la séquence manuelle concernée (Baker-Schenk, 1983). Mais dès lors que l'on observe des phénomènes d'alignement entre mouvements non manuels et la structure prosodique du discours dans les langues vocales (Graf et al., 2002, par exemple), les distinctions [intonation vocale = non-manuel en langue des signes] et [segmental = manuel] / [suprasegmental = non-manuel] nous paraissent inopérantes.

1.1.2 *Le repérage de motifs prosodiques*

Dans les analyses prosodiques de la langue des signes américaine notamment, nous retrouvons la notion de durée (des mouvements, des tenues) qui concerne plus directement le manuel (Grosjean, 1979, par exemple). Mais est également présente la notion d'accentuation ou d'emphase qui associe articulateurs manuels et non manuels (cf. Wilbur, 1999 pour une synthèse). Boyes-Braem (1999) s'est intéressée aux mouvements de translation latérale du torse dans des énoncés de signeurs natifs en langue des signes de Suisse allemande. L'auteur conclut que ces balancements coïncident en partie avec des frontières de propositions, et, à une plus grande échelle, à des unités de discours. En création, les paramètres prosodiques constituent des ressources poétiques dans la structuration contrôlée du discours. Ainsi les auteurs peuvent choisir de respecter un tempo isochrone (avec des battements à intervalle de temps régulier), une construction spatiale privilégiant la symétrie ou l'équilibre des zones de l'espace grâce à la répartition des articulateurs main droite, main gauche, par exemple (cf. Blondel & Miller, 2009).

La traduction d'un registre poétique constitue alors un véritable défi, notamment pour la transmission la plus fidèle possible de l'intention du poète, incluant son architecture prosodique et les effets ainsi provoqués sur le public.

1.2 Le recours à la capture de mouvements pour l'étude des langues des signes

Les études 'phonétiques' (au sens de description articulatoire) mentionnées ont été enrichies depuis quelques années par le recours à la capture de mouvements (mocap) et à des dispositifs de moins en moins invasifs. Ce qui était auparavant mesuré sur un support vidéo peut désormais être validé, complété et précisé par des mesures biomécaniques. Ainsi, Tyrone et al. (2010) étudient l'élongation relative du mouvement d'un signe (en langue des signes américaine) selon sa position (initiale, médiane ou finale) dans la proposition et constatent que, en contexte final, la phase de repos (*release*) du mouvement du signe est allongée proportionnellement. De nouvelles pistes sont également ouvertes, notamment en s'appuyant sur l'élicitation et le jugement de perception, en attendant qu'une détection automatique efficace soit mise au point. Ainsi, Tanaka & van der Hulst (2004) observent, lors d'une tâche d'élicitation d'emphase en langue des signes japonaise, que des signeurs font varier non seulement les paramètres du mouvement de chaque signe, mais également ceux du mouvement de transition entre les signes. Jantunen (2003) a également recours à la mocap et utilise la vitesse et l'accélération pour comparer les mouvements des signes et les mouvements des transitions entre les signes, en langue des signes finlandaise. Il observe principalement que la vitesse des signes est inférieure à celle des transitions mais que les signes présentent davantage de phénomènes d'accélération (et de variation) que les transitions.

2 Contexte de l'étude, expérimentations, données

En lien avec cette complémentarité d'informations entre la vitesse et ses dérivées, Wilbur & Martinez (2002) ont effectué une analyse comparée de la perception respective des variations des paramètres de vitesse, d'accélération et de saccade (ou *jerk*). Les auteurs cherchaient à comprendre si l'un de ces paramètres cinématiques pouvait, dans l'encodage prosodique de la langue des signes américaine, servir plus efficacement de marqueur que les deux autres. Ils s'appuient sur la capture des mouvements des bras et mains, postulant que les mouvements ont des répercussions sur les bras, même s'ils ont été initiés par une autre partie du corps. Ce set de capture est proche de celui mis en place dans le programme de recherche CIGALE, projet pluridisciplinaire, scientifique et artistique, d'étude du mouvement gestuel (Batras et al., 2015). L'expérimentation présentée ici est née de la confrontation entre la pratique de l'interprétation en langue des signes (Catteau, 2015) et ce programme de recherche.

2.1 CIGALE et les données poétiques

Depuis 2013, l'équipe du projet CIGALE a enregistré un corpus regroupant quatre types de données gestuelles (gestualité coverbale, mime, direction de chœur, langue des signes poétique). Chaque sous-corpus a été enregistré grâce à un système de capture de mouvement 3D composé de vingt-quatre caméras numériques infrarouges (VICON, 120 fps), qui enregistrent les positions successives de plus de quatre-vingt-dix marqueurs réfléchissants placés sur le sujet selon des positions anatomiques précises. Ce *marker set* suit une méthode standardisée pour que chaque segment du corps apparaisse comme étant un solide indéformable. Il permet un changement de repère à même d'isoler le mouvement de chaque segment (Dumas et al., 2012).

Les séquences de langue des signes poétique ont été préparées et réalisées par Jules Turlet, poète et chansigneur[2] sourd. Six séquences poétiques y sont déclinées avec des variations prosodiques (des modifications de la vitesse ou de l'amplitude du mouvement principalement). Le corpus que nous avons constitué pour la présente étude est composé de trois versions de trois séquences poétiques dénommées « Arbre », « Rivière » et « Sourire ». Chacune de ces trois versions est une proposition du même contenu segmental, avec une modification du débit entraînant donc une variation suprasegmentale. Pour modifier le débit de sa performance, le signeur a utilisé divers procédés linguistiques tel un ajustement de l'amplitude des signes ou la réduction des temps de pause[3].

2.2 Les traductions LSF / français vocal

Outre la capture de mouvement, les séquences de langue des signes poétique ont été filmées à l'aide d'une caméra ordinaire. En vue de leur traduction vocale, nous avons organisé une rencontre avec six interprètes en langue des signes (experts en traduction de registre poétique). Chaque interprète a reçu les séquences « Arbre », « Rivière » et « Sourire » incluant leurs trois variantes rythmiques, soit neuf courtes séquences poétiques à traduire. Nous avons effectué un enregistrement audio et audio-vidéo des traductions, notamment avec le support vidéo de l'extrait poétique, contraignant ainsi l'interprète à traduire en respectant la prosodie du signeur.

[2] Le chansigne est la traduction et adaptation d'une chanson en langue des signes ou une création directe en langues des signes qui se caractérise par un tempo régulier.

[3] Voir Catteau (2015 : 32-33), pour une description et une analyse complètes de la gestion de la variation rythmique des séquences.

2.3 Méthodologie

Afin d'identifier les stratégies interprétatives face à la variation rythmique des différents stimuli, nous avons effectué un premier repérage qualitatif sur vidéo (comme la durée des pauses ou les mouvements d'impulsion) et exploité ensuite des mesures quantitatives extraites de la capture de mouvement (pour le signeur, cf. Catteau, 2015 : 27-34) et de l'enregistrement audio (pour les traductions en français vocal, cf. Catteau, 2015 : 34-51).

2.3.1 *Annotation manuelle et mesures des fichiers audio*

Dans un premier temps, nous avons annoté manuellement les cinquante-deux traductions recueillies en simultané à la vidéo, en ciblant les éléments prosodiques saillants. Nous avons pour cela effectué une segmentation syllabique des traductions vocales, en codant les syllabes courtes et longues, selon la durée relative entre l'attaque d'une syllabe et l'attaque de la syllabe suivante suivante (avec [.] pour les courtes et [_] pour les longues). Nous avons noté les respirations (avec le symbole [V] emprunté à la notation musicale), les pauses (avec une croix cerclée [⊗]) et chaque impulsion vocale ([↑]), comme l'illustre la figure 1.

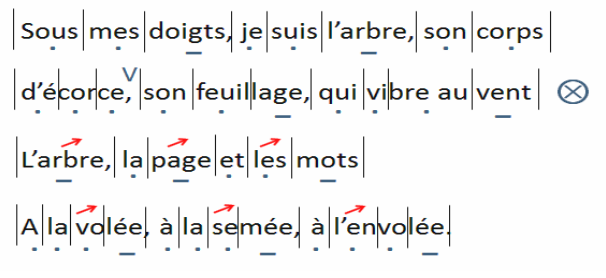


FIGURE 1 : Exemple d'annotation d'une séquence de traduction vocale

Dans un second temps, grâce au logiciel PRAAT[4], nous avons analysé les formes d'onde des traductions vocales. Nous avons calculé la durée des séquences de traduction vocale, ainsi que la durée des pauses dans le flux vocal des interprètes.

2.3.2 *Mesures biomécaniques*

Des mesures biomécaniques ont été ajoutées à l'analyse afin de mesurer de manière objective les variations prosodiques du signeur. Parmi les différentes mesures, nous avons utilisé la quantité de mouvement, précédemment employée pour caractériser la dynamique des gestes réalisés pour diriger un orchestre et provoquer différents types de nuances (Sarasua & Guaus, 2014). Elle correspond à la moyenne des vitesses de toutes les jointures du corps. Elle permet de caractériser l'intensité globale des mouvements effectués et d'appréhender les variations de prosodie.

Nous exploitons également les matrices de similarité (Foote, 1999), une approche issue de l'analyse musicale, utilisée pour structurer automatiquement des pièces sonores. Elles sont employées pour visualiser de manière compacte les formes répétitives d'un ensemble de descripteurs temporels. Appliquées aux positions des marqueurs, elles permettent d'appréhender la prosodie en termes de

[4] http://www.fon.hum.uva.nl/praat/

répétition et de transition. Les positions spatiales des marqueurs ont dans ce cas été normalisées par la position relative et les rotations du corps du signeur. Enfin, pour modéliser certains mouvements de transition du signeur, nous avons utilisé les rotations d'un axe horizontal passant par les hanches.

3 Résultats

Les séquences poétiques se caractérisent par des motifs rythmiques et spatio-mélodiques qui jouent sur les variations de durée, d'amplitude et de vélocité. Les traductions proposées en français vocal exploitent ces variations, tout en présentant des variantes dans les stratégies employées. Nous présentons l'ensemble des observations concernant la durée des unités et celle des tenues ou pauses (3.1.1), les observations concernant les impulsions et leur rôle dans l'élaboration de motifs répétés (3.1.2), et enfin l'examen des mouvements de transition (3.1.3).

3.1.1 *Durées, pauses et tenues*

Le signeur a fait varier respectivement la durée de chaque séquence en allongeant ou rétrécissant la durée de ses mouvements, notamment en leur donnant plus ou moins d'amplitude. Nous nous attendons à ce que le débit vocal des interprètes soit modifié dans l'exercice de traduction des différentes versions.

Les interprètes ont utilisé trois types de stratégies interprétatives pour faire varier la durée totale des traductions (cf. Catteau, 2015:43-51) : premièrement, en jouant sur la durée des syllabes ; deuxièmement, en modifiant le texte d'une variation rythmique à l'autre (plus de mots en version lente qu'en version normale et inversement dans la version rapide *versus* normale) ; troisièmement, en faisant varier la durée des pauses et la place des respirations. La figure 2 illustre ainsi des longueurs différentes d'une même pause effectuée dans la traduction du poème « Arbre », en fonction des versions « normale », « lente » ou « rapide » de la performance en LSF. Les interprètes qui ont fait varier aussi le texte pour chaque variante rythmique ont effectué plus de pauses en version lente qu'en normale (et donc de temps de pauses sur l'ensemble du poème) et inversement en version rapide.

FIGURE 2 : Exemple de forme d'onde pour une séquence de traduction vocale[5]. La durée des pauses varie en fonction de la vitesse d'exécution de la performance en LSF.

3.1.2 *Impulsions et motifs répétés*

Outre l'annotation des traductions vocales (cf. 2.3.1), nous avons procédé à une annotation manuelle des séquences poétiques en langue des signes à l'aide du logiciel ELAN[6]. Nous avons notamment

[5] Traduction : « Crépusculaire rivière. De tes eaux naissent et meurent les bois, abattus, engloutis, submergés par tes flots. Crépusculaire, dévorante rivière. » .
[6] https://tla.mpi.nl/tools/tla-tools/elan/

découpé chaque séquence en phrases poétiques, commencé une description articulatoire du mouvement et nous avons ainsi repéré les impulsions gestuelles marquantes de ces phrases.

Les courbes cinématiques de la quantité de mouvement font apparaître des pics d'accélération suivis d'une brusque décélération. Ces pics correspondent aux impulsions annotées manuellement dans les séquences poétiques. Par ailleurs, nous avons remarqué que les motifs répétés, précédemment repérés lors de nos annotations manuelles, apparaissaient bien sur ces courbes (voir la figure 3).

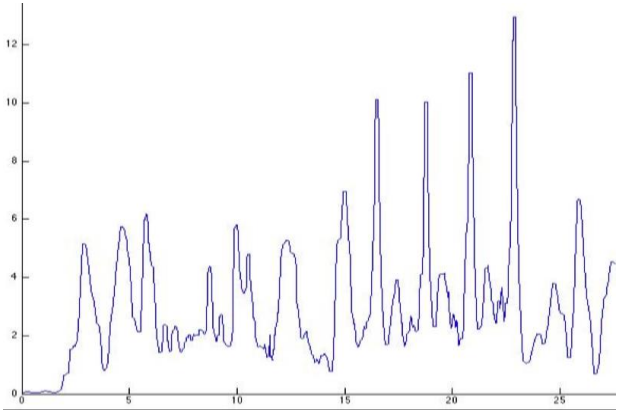


FIGURE 3 : Courbe d'intensité de quantité de mouvement en fonction du temps, calculé à partir du poème "Arbre". Les motifs répétés apparaissent des secondes 15 à 23.

La répétition de mouvements apparaît également clairement sur les matrices de similarité calculées sur la performance signée. La figure 4 illustre les similarités entre les positions spatiales des marqueurs normalisées par rapport à la position des hanches. Les distances entre chaque couple de positions sont calculées et illustrées par des tons allant du bleu (distance importante, faible similarité) au rouge (faible distance, forte similarité). Elle permet d'observer que le geste poétique se découpe en trois phases différentes. Ces dernières se matérialisent par des zones homogènes, dont le découpage correspond aux respirations identifiées dans la traduction vocale. Des répétitions sont également visibles sur la dernière partie, où quatre mouvements similaires se succèdent.

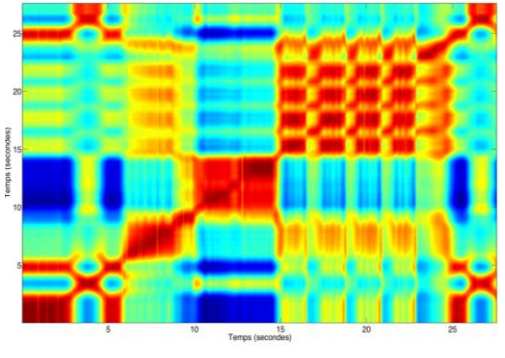


FIGURE 4 : Matrice de similarité calculée sur les distances entre jointures lors de la réalisation du poème "Arbre". Le geste poétique (réalisé de 5 à 25 secondes) se découpe en trois phases (6-10, 10-15, 15-23). Les mouvements répétés sont visibles sur la partie en damier (15-23).

3.1.3 *Focus sur les mouvements de transition*

La matrice de similarité permet aussi d'illustrer des mouvements de transition. Par exemple, le segment allant de 10 à 15 secondes de la figure 4 est assez éloigné du reste de la performance, car il est constitué par une position du corps différente. Par ailleurs, les mouvements de transition liés à la rotation des hanches (figure 5) permettent d'obtenir une autre lecture de la performance s'appuyant sur deux grands segments entrecoupés d'un mouvement de transition (secondes 14 à 15).

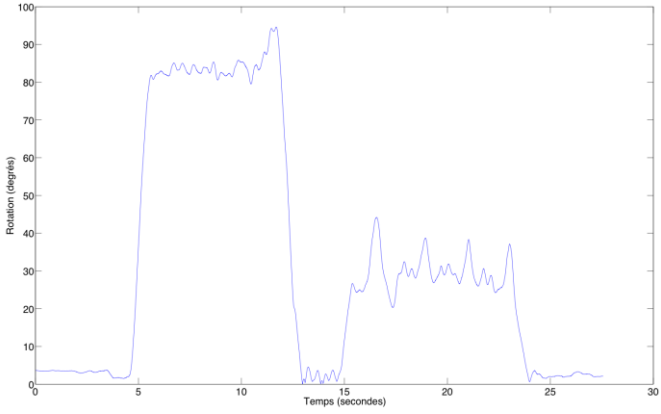


FIGURE 5 : Rotation de l'axe des hanches en fonction du temps, lors de la réalisation du poème "Arbre".

4 Pour ne pas conclure

Dans le cadre de notre recherche des points communs et des spécificités des prosodies vocale et gestuelle, nous avons étudié les variations de durée, fréquence et intensité. Il nous semble pertinent d'explorer la durée des séquences de flux gestuel comme celles de flux sonore, en intégrant la dimension des pauses. Ces mesures concernant la durée des gestes (vocaux ou corporels) ont un lien avec les paramètres d'intensité-amplitude d'une part et avec les paramètres de fréquence-vélocité (et dérivées) d'autre part. Les mesures de ces dernières nécessitent le recours à la mocap et gagnent ainsi en finesse de granularité.

Nous pouvons distinguer des *motifs* ou contours prosodiques systématiques en nous appuyant sur les régularités et contrastes dessinés par la variation de ces paramètres, comme par exemple les pics de quantité de mouvement correspondant à des impulsions. A l'avenir, cette technique pourrait permettre d'annoter automatiquement de nouveaux motifs. De la même façon, les matrices de similarité permettent de structurer objectivement une performance en LSF.

Cette étude peut être exploitée pour l'explicitation des stratégies de traduction-interprétation, *a fortiori* dans le passage de la LSF vers le français vocal, qui semble un défi singulier pour les apprenants. La formation d'interprète français-LSF se focalise en effet sur la déverbalisation (technique de traduction se détachant des structures proprement linguistiques pour accéder au sens), le lexique, les prises de rôle (niveau énonciatif), les placements (niveau morphosyntaxique) et s'intéresse peu au rôle de la prosodie (tant vers la LSF que vers le français). La poursuite des recherches énoncées dans cet article pourrait, à l'avenir, aider à la prise de conscience de l'importance du contenu suprasegmental.

Remerciements

Nous remercions tous nos partenaires du projet CIGALE, Jules Turlet pour sa contribution poétique en LSF et les interprètes qui ont participé à l'expérimentation : Vincent Bexiga, Alexandra Bilisko, Carlos Carreras, Aurore Corominas, Marie Lamothe et Emile Tolian.

Références

BAKER-SCHENK C. L. (1983). *A Microanalysis of the Nonmanual Components of Questions in American Sign Language.* Unpublished dissertation, University of California, Berkeley, CA.

BATRAS D., BLONDEL M., BOUTET D., CHEN C.-Y., CATTEAU F., GUEZ J., GUYOT P., JÉGO J.-F., TRAMUS M.-H., VINCENT C. (2015). On the CIGALE project. In *The Digital Subject: Codes*, Paris.

BLONDEL M., LE GAC D. (2007). Entre parenthèses y a-t-il une intonation en LSF ? *Silexicales* 1-16.

BLONDEL M., MILLER C. (2009). Symmetry and children's poetry in sign languages. Dans J.-L. Aroui & A. Arleo (eds.), *Towards a typology of poetic forms, Language Faculty and Beyond*, 2. Amsterdam : John Benjamins PC, 143-163.

BOYES-BRAEM P. (1999). Rhythmic temporal patterns in the signing of deaf early and late learners of Swiss German Sign Language. *Language and Speech*, 42(2-3), 177-208.

CATTEAU F. (2015). *La prosodie de l'interprète en traduction de poèmes signés : quelles stratégies mises en place face à la variation rythmique ?* Mémoire de recherche, Master Interprétariat en langue des signes, Université Paris 8.

COROMINAS A. (2015). *La traduction française de poésie créée en LSF*, Mémoire de recherche, Master Interprétariat en langue des signes, Université Paris 8.

DUMAS R., ROBERT T., POMERO V., CHEZE L. (2012). Joint and segment coordinate systems revisited. *Computer methods in biomechanics and biomedical engineering* 15 (suppl. 1), 183-185.

FERRÉ G. (2014). Tension gestuelle en co-présence d'un accent d'intensité. Actes des *Journées d'Etude sur la Parole (JEP)*, Le Mans [CD-Rom].

FOOTE J. (1999). Visualizing music and audio using self-similarity. *Proceedings of the 7th ACM international conference on Multimedia (Part 1)*, 77-80.

GRAF H. P., COSATTO E., STROM V., HUANG F. J. (2002). Visual Prosody: Facial Movements Accompanying Speech. *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*.

GROSJEAN F. (1979). A study of timing in a manual and a spoken language: American Sign Language and English. *Journal of Psycholinguistic Research* 8(4), 379-405.

JANTUNEN T. (2013). Signs and transitions: Do they differ phonetically and does it matter? *Sign Language Studies* 13(2), 211-237.

SARASÚA Á., GUAUS E. (2014). Dynamics in Music Conducting: A Computational Comparative Study Among Subjects. Actes de *NIME*, 195-200.

SANDLER W. (1999). Prosody in two natural language modalities. *Language and Speech* 42 (2-3), 127-142.

TANAKA S., VAN DER HULST H. (2004). Speed of hand movement: a quantitative study. Poster à *TISLR 8*, Barcelona.

TYRONE M. E., NAM H., SALTZMAN E., MATHUR G., GOLDSTEIN L. (2010). Prosody and Movement in American Sign Language: A Task-Dynamics Approach. Actes de *Speech Prosody 2010*, Chicago.

WILBUR R. B., MARTINEZ A. (2002). Physical correlates of prosodic structure in American Sign Language. In M. Andronis, E. Debenport, A. Pycha & K. Yoshimura (eds.), *CLS* 38, 693-704.

Voix de femmes, voix d'hommes: une étude du voice onset time, de la répartition consonnes/voyelles et du débit de parole chez des locuteurs francophones et anglophones américains

Erwan Pépiot

Groupe LECSeL – EA1569
Université Paris 8 - 2 rue de la liberté 93200 Saint-Denis
erwan.pepiot@free.fr

RÉSUMÉ

La présente étude est une analyse acoustique de mots et pseudo-mots de type /CVCV/ produits par des locuteurs anglophones du nord-est des États-Unis (5 femmes, 5 hommes) et des francophones parisiens (5 femmes, 5 hommes). Le VOT des consonnes occlusives initiales, la durée des énoncés, ainsi que la répartition temporelle consonnes/voyelles ont été mesurés. Des différences inter-genres significatives ont été observées dans les deux langues sur chacun des paramètres testés : le contraste de VOT entre les occlusives sourdes et voisées s'est révélé plus important chez les locutrices, le débit de parole plus élevé chez les locuteurs masculins, et la proportion occupée par les consonnes plus importantes chez les femmes. Ces résultats suggèrent une tendance à la recherche d'une plus grande intelligibilité chez les locutrices. Les différences acoustiques femmes-hommes seraient donc en partie construites socialement.

ABSTRACT

Female and male speech: a study of VOT, C/V temporal distribution and speech rate in Parisian French and American English speakers

The current study is an acoustic analysis of /CVCV/ disyllabic words produced by 10 Parisian French speakers (5 females, 5 males) and 10 Northeastern American English speakers (5 females, 5 males). Voice onset time in initial stop consonants was measured, as well as words' duration and consonants/vowels temporal distribution. Significant cross-gender differences were obtained for each tested parameter: VOT contrasts between voiced and voiceless consonants were larger in female speakers; speech rate was faster in male speakers; while the proportion of consonants within the words was greater in female speakers. Such results support the idea that female speakers would try to achieve greater intelligibility than males and suggest that cross-gender acoustic differences are partly socially constructed.

MOTS-CLÉS : voix de femmes, voix d'hommes, différences inter-genres, VOT, vitesse d'élocution.

KEYWORDS: female voices, male voices, cross-gender differences, VOT, speech rate.

1 Introduction

La fréquence fondamentale moyenne est communément considérée comme la principale différence entre les voix de femmes et d'hommes. Elle se situerait autour des 120 Hz chez les locuteurs masculins, contre environ 200 Hz chez les locutrices (Boë et al., 1975). Néanmoins, plusieurs études ont révélé d'autres différences inter-genres.

Tout d'abord, les formants vocaliques des locutrices se situent généralement dans des fréquences plus élevées que leurs homologues masculins. L'ampleur de ces différences varie d'une étude à l'autre et semble dépendre du formant et du type de voyelle (Whiteside, 2001). Une tendance similaire a été observée sur les consonnes, dont le bruit se trouve renforcé à des fréquences plus hautes chez les locutrices (Fox & Nissen, 2005).

Concernant les différences hommes-femmes dans le domaine temporel, plusieurs auteurs se sont penchés sur le voice onset time (VOT). Chez les locuteurs anglophones, et bien que les résultats varient sensiblement d'une étude à l'autre, la tendance générale irait vers un délai d'établissement du voisement plus long chez les femmes. A titre d'exemple, Swartz (1992) a mis en évidence des VOT significativement plus courts chez les hommes anglophones, sur la paire de consonnes occlusives alvéolaires [t] / [d]. Des résultats similaires ont été obtenus à plusieurs reprises sur les occlusives sourdes, toujours sur des locuteurs anglophones (Whiteside & Irving, 1997 ; Robb et al., 2005).

Cependant, l'ensemble des études mentionnées ici ont été menées sur des locuteurs anglophones. L'analyse de données complémentaires provenant d'autres langues est donc plus que nécessaire. Une étude plus récente montre d'ailleurs que chez les locuteurs du coréen, la tendance semble opposée à celle observée chez les anglophones : les VOT des locutrices seraient, selon le type de consonne, soit significativement plus courts que ceux des hommes, soit de durée équivalente (Oh, 2011). De même, une étude conduite sur des locuteurs suédophones (Karlsson et al., 2004) ne fait état d'aucune différence significative entre les VOT des hommes et des femmes adultes.

La durée des énoncés a elle aussi fait l'objet d'études comparatives entre les deux genres. Dans une étude portant sur plus de 600 locuteurs de l'anglais américain (Byrd, 1994), les énoncés produits par des hommes présentent une durée moyenne de 6,2 % inférieure à ceux produits par des femmes. Une tendance analogue a été observée chez des locuteurs de l'anglais britannique ou irlandais (Whiteside, 1995 ; Fitzsimons et al., 2001). Néanmoins, plusieurs autres études n'ont mis en évidence aucune différence inter-genres significative sur ce paramètre (Jacewicz et al., 2009 ; Simpson et Ericsdotter, 2003). La répartition consonnes-voyelles au sein des énoncés, n'a quant à elle fait l'objet que de très peu d'attention, sur le plan de la comparaison femmes-hommes.

Une partie de ces variations acoustiques inter-genres trouveraient leur origine dans des différences anatomiques et physiologiques qui émergent à la puberté. Les plis vocaux des sujets adultes sont sensiblement plus longs et plus épais chez les locuteurs masculins (Kahane, 1978), ce qui expliquerait les différences observées sur la fréquence fondamentale. L'autre élément important concerne la longueur du conduit vocal. Ce dernier mesure en moyenne 14.5 cm chez les femmes et 17 à 18 cm chez les hommes adultes (Simpson, 2009). Ces caractéristiques permettent de rendre compte, au moins partiellement, des différences inter-genres observées sur les formants vocaliques et les zones de bruits des consonnes.

En revanche, d'autres différences inter-genres, comme celles observées sur la durée des énoncés, semblent ne pouvoir s'expliquer que par des facteurs culturels. De plus, certaines de ces différences acoustiques pourraient fortement varier en fonction de la langue du locuteur. C'est en tout cas ce que suggère une étude menée par Johnson (2006) sur les formants vocaliques. Cependant, il ne s'agit ici que d'une collection de données obtenues dans différentes études expérimentales, conduites par des auteurs différents et avec des protocoles expérimentaux distincts. D'autre part,

aucune investigation similaire n'a été menée sur les autres paramètres acoustiques variant en fonction du genre du locuteur.

Par conséquent, il semble particulièrement pertinent de conduire une étude expérimentale inter-langues sur les différences acoustiques femmes-hommes. Dans la présente étude, les différences inter-genres sur le plan du VOT, du débit de parole et de la répartition temporelle consonnes-voyelles, ont été analysées chez des locuteurs anglophones américains et francophones parisiens. L'hypothèse générale étant la suivante : *il existe des différences inter-genres significatives sur chacun de ces paramètres acoustiques*.

2 Méthode

2.1 Matériau linguistique

Du matériau linguistique anglais et français était nécessaire pour réaliser cette étude. Des mots et pseudo-mots dissyllabiques ont été utilisés, afin de permettre de tester différentes combinaisons de phonèmes. Leur sélection a été réalisée sur la base de deux critères principaux : rendre les deux corpus le plus similaires possible (i.e. utilisation de segments qui, en anglais et en français, présentent des qualités relativement proches), tout en limitant le nombre de combinaisons en choisissant uniquement les phonèmes les plus pertinents. Pour ce faire, la dernière séquence CV était identique sur chacun des items : c'est la combinaison /pi/ qui a été préférée, car elle peut apparaître en fin de mot aussi bien en anglais qu'en français. De plus, certaines consonnes initiales, telles que les labiales [p] / [b] et [f] / [v] ont été écartées. Vingt-sept mots /(C)Vpi/ ont ainsi été retenus pour chaque langue :

- Combinaisons /C (occlusive) – V – p – i/ : /tipi/, /tapi/, /tupi/, /dipi/, /dapi/, /dupi/, /kipi/, /kapi/, /kupi/, /gipi/, /gapi/, /gupi/ pour le corpus français, /ˈti:pi/ , /ˈtæpi/, /ˈtu:pi/, /ˈdi:pi/, /ˈdæpi/, /ˈdu:pi/, /ˈki:pi/, /ˈkæpi/, /ˈku:pi/, /ˈgi:pi/, /ˈgæpi/, /ˈgu:pi/ pour le corpus anglais.
- Combinaisons /C (fricative) – V – p – i/ : /sipi/, /sapi/, /supi/, /zipi/, /zapi/, /zupi/, /ʃipi/, /ʃapi/, /ʃupi/, /ʒipi/, /ʒapi/, /ʒupi/ pour le corpus français, /ˈsi:pi/, /ˈsæpi/, /ˈsu:pi/, /ˈzi:pi/, /ˈzæpi/, /ˈzu:pi/, /ˈʃi:pi/, /ˈʃæpi/, /ˈʃu:pi/, /ˈʒi:pi/, /ˈʒæpi/, /ˈʒu:pi/ pour le corpus anglais.
- Combinaisons /V – p – i / : /ipi/, /api/, /upi/ pour le corpus français, /ˈi:pi/, /ˈæpi/, /ˈu:pi/ pour le corpus anglais.

Il n'y a pas d'accent lexical en français (Di Cristo, 1999), mais à l'intérieur de la phrase porteuse utilisée pour les enregistrements (voir 2.3), les locuteurs francophones ont naturellement produit un léger accent emphatique sur la première syllabe des mots expérimentaux.

2.2 Participants

Vingt locuteurs monolingues ont été enregistrés. Dix d'entre eux sont des locuteurs francophones natifs (5 femmes, 5 hommes), les dix autres des locuteurs anglophones américains natifs (5 femmes, 5 hommes). Les 10 locuteurs américains sont tous originaires d'une même région du Nord-Est des États-Unis (États de Pennsylvanie, du Massachussetts, de New-York et du Vermont). Les 10 locuteurs francophones sont quant à eux originaires de la région parisienne. Tous les locuteurs étaient âgés de 20 à 40 ans (*SD*=6,5 ans) lors des enregistrements. La moyenne d'âge était de 28.2 ans pour les anglophones américains (29,4 chez les femmes, 27 chez les hommes) et de 26,6 ans pour les francophones (27,2 chez les femmes, 26 chez les hommes).

Tous les locuteurs sont non-fumeurs, et ne présentent pas de troubles de la parole. Chacun d'entre eux a reçu une clé USB en échange de sa participation à la présente étude.

2.3 Procédure d'enregistrement

Les enregistrements se sont déroulés dans une chambre anéchoïque. L'enregistreur numérique utilisé est un *Edirol R09-HR* de marque *Roland*. Les locuteurs anglophones ont eut à lire le corpus anglais, les francophones le corpus français. Les mots ont été présentés orthographiquement aux participants. Afin de maintenir les paramètres prosodiques constants, les mots dissyllabiques ont été placés dans une phrase porteuse : « Il a dit '*MOT'* deux fois » pour le corpus français, « He said '*WORD'* three times » pour le corpus anglais. Il a été demandé aux locuteurs de lire chaque phrase deux fois, avec un débit de parole normal.

2.4 Analyse des données

Les mots ont dans un premier temps été extraits de la phrase cadre. Tous les items ayant été enregistrés deux fois, c'est l'occurrence la plus satisfaisante acoustiquement de chacun d'entre eux qui a été retenue (absence de bruit parasite, d'hésitations de la part du locuteur…). Le total des items expérimentaux s'élève donc à 270 pour chaque langue (27 mots * 10 locuteurs). Les mots ont ensuite été segmentés en phones puis étiquetés. L'ensemble de ces opérations ont été réalisées dans le logiciel *Praat*.

Le *voice onset time* a été mesuré manuellement pour chaque consonne occlusive initiale, à partir du spectrogramme. Le moment du relâchement de l'occlusive ainsi que le début du voisement ont donc dû être localisés, le VOT correspondant à l'espacement temporel entre le premier repère et le second (si le voisement débute avant le relâchement de la consonne, le VOT aura une valeur négative, comme pour les occlusives voisées du français). Il est important d'évoquer ici le cas particulier des consonnes occlusives de l'anglais : lorsqu'elles se trouvent en position initiale de mot, devant une voyelle et sous l'accent lexical, les consonnes /t/ et /k/ se réalisent au niveau phonétique [t^h] et [k^h] (occlusives sourdes aspirées), tandis que les phonèmes /d/ et /g/ sont généralement produits comme des occlusives dévoisées non-aspirées [d̥] et [g̥], le voisement débutant uniquement au début de la voyelle suivante (voir notamment Lin & Wong, 2011).

La durée des mots dissyllabiques a également été recueillie *via* le logiciel Praat, de même que la durée de chacune des consonnes et voyelles composant ces mots.

Dans un deuxième temps, afin de tester si les différences inter-genres observées sont significatives, des ANOVAs ont été conduites sur les données recueillies, dans les deux langues et sur chaque paramètre acoustique.

3 Résultats

3.1 Voice onset time

La durée moyenne du VOT des consonnes occlusives produites par les femmes et les hommes francophones est visible dans la Table 1, ci-après.

On constate que le VOT est en moyenne nettement plus élevé chez les locutrices sur les occlusives sourdes (+40%). C'est en revanche le contraire pour les occlusives voisées : le VOT est globalement plus court chez les femmes que chez les hommes francophones, de l'ordre de 33%. Afin de vérifier si ces différences sont significatives, j'ai conduit deux ANOVAs à deux facteurs (« genre du locuteur » et « consonne »). L'une porte sur les VOT des occlusives sourdes, l'autre sur ceux des occlusives voisées. Dans les deux cas, les tests concluent à l'existence d'une différence inter-genre forte et significative : $F_{(1,56)}$=9,047 ; p=0,0039 pour les consonnes sourdes, $F_{(1,56)}$=39,917 ; p<0,0001 pour les consonnes sonores.

	VOT moyen (ms)		
Consonne	**Femmes**	**Hommes**	**Ratio F/H**
[t]	**51**	**36**	**1,43**
[k]	**62**	**45**	**1,37**
Occlusives sourdes	**56**	**40**	**1,40**
[d]	**-95**	**-65**	**1,46**
[g]	**-76**	**-63**	**1,21**
Occlusives voisées	**-86**	**-64**	**1,33**

TABLE 1 – VOT moyen (ms) des consonnes occlusives initiales sourdes et voisées pour les femmes et les hommes francophones, sur un total de 15 mesures par consonne et par genre (5 locuteurs * 3 items). Le ratio femmes-hommes est également indiqué pour chaque consonne et type de consonne.

Le VOT moyen des occlusives produites par les locuteurs et locutrices anglophones est présenté dans la Table 2, ci-dessous.

Chez les anglophones américains, le VOT est fortement plus élevé chez les locutrices que chez les locuteurs sur les occlusives aspirées (+42%). Concernant les occlusives non-aspirées, on constate également un VOT supérieur chez les femmes, mais dans des proportions bien moindres (+11%).

	VOT moyen (ms)		
Consonne	**Femmes**	**Hommes**	**Ratio F/H**
[t^h]	**83**	**59**	**1,41**
[k^h]	**92**	**64**	**1,44**
Occlusives aspirées	**87**	**61**	**1,42**
[d̥]	**22**	**17**	**1,24**
[g̥]	**26**	**26**	**1,02**
Occlusives non-aspirées	**24**	**22**	**1,11**

TABLE 2 – VOT moyen (ms) des consonnes occlusives initiales aspirées et non-aspirées pour les femmes et les hommes anglophones américains, sur un total de 15 mesures par consonne et par genre (5 locuteurs * 3 items). Le ratio femmes / hommes est également indiqué pour chaque consonne et type de consonne.

Une ANOVA à deux facteurs (« genre du locuteur » et « consonne »), indique que la différence femmes-hommes est très forte et largement significative sur les occlusives aspirées : $F_{(1,56)}$=62,031 ; p<0,0001. A l'inverse, la même analyse effectuée sur les données relatives aux occlusives non-aspirées, montre que la différence inter-genre n'atteint pas de seuil de significativité pour ce type de consonnes avec $F_{(1,56)}$=3,378 et p=0,0714.

Le contraste moyen (en terme de VOT) entre les consonnes sourdes et voisées, ou aspirées et non-aspirées, a également été calculé. Les données relatives aux locutrices et locuteurs francophones sont disponibles ci-dessous (Table 3).

Chez les locuteurs francophones, le contraste de VOT entre occlusives sourdes et voisées est donc nettement plus accentué chez les locutrices, et ce quelle que soit la paire de consonne. Toutes consonnes confondues, la différence de VOT entre les occlusives sourdes et voisées est en moyenne 35% plus élevée chez les femmes.

	Contraste de VOT moyen (ms)		
Consonnes	**Femmes**	**Hommes**	**Ratio F/H**
[t] vs. [d]	**146**	**101**	**1,45**
[k] vs. [g]	**138**	**108**	**1,28**
Sourdes Vs. voisées	**142**	**105**	**1,35**

TABLE 3 – Contraste de VOT moyen (ms) sur les paires de consonnes sourdes / voisées pour les femmes et les hommes francophones. Le ratio femmes / hommes est également indiqué pour chaque paire de consonnes.

Une ANOVA à un facteur (« genre du locuteur ») effectuée sur ces données confirme que la différence est largement significative avec $F_{(1,58)}$=60,332 et p<0.0001.

Le contraste de VOT moyen pour les anglophones américains est présenté dans la Table 4, ci-après. Une tendance similaire apparaît donc pour les locuteurs anglophones : malgré l'existence d'un VOT plus élevé chez les locutrices sur les deux types de consonnes, le contraste entre occlusives aspirées et non-aspirées demeure au final globalement plus marqué chez les femmes que chez les hommes d'environ 60%.

	Contraste de VOT moyen (ms)		
Consonnes	**Femmes**	**Hommes**	**Ratio F/H**
[t^h] vs. [ḍ]	**61**	**42**	**1,47**
[k^h] vs. [g̰]	**66**	**38**	**1,73**
Aspirées Vs non-aspirées	**64**	**40**	**1,60**

TABLE 4 – Contraste de VOT moyen (ms) sur les paires de consonnes aspirées / non-aspirées pour les femmes et les hommes anglophones américains. Le ratio femmes / hommes est également indiqué pour chaque paire de consonnes.

Un test statistique identique à celui conduit sur les données des francophones confirme que cette différence est très nettement significative : $F_{(1,58)}$=58,902 ; p<0.0001.

3.2 Durée des mots

La durée moyenne de mots (ms) produits par les locuteurs francophones et anglophones en fonction du genre du participant est présentée dans le tableau 5, ci-dessous.

	Locuteurs francophones			**Locuteurs anglophones**		
	Femmes	**Hommes**	**Ratio F/H**	**Femmes**	**Hommes**	**Ratio F/H**
Durée moyenne des mots (ms)	**510**	**445**	**1,15**	**555**	**441**	**1,26**
SD	*90*	*58*		*77*	*54*	

TABLE 5 – Durée moyenne (ms) des 27 mots dissyllabiques de type (C)VCV pour les femmes francophones (n=5), les hommes francophones (n=5), les femmes anglophones (n=5) et les hommes anglophones (n=5). L'écart type (SD) sur les 135 items (27 mots * 5 locuteurs) est également indiqué pour chacun des quatre groupes.

Les résultats indiquent que la durée moyenne des mots est plus importante chez les locutrices dans les deux langues. Cette différence inter-genres est plus marquée chez les anglophones

américains (+26%) que chez les francophones (+15%). Les mots produits par les locuteurs masculins des deux langues étant d'une longueur équivalente, cette variation inter-langues s'explique par la durée des items produits par les locutrices, qui est sensiblement plus élevée chez les anglophones (+9%).

Une ANOVA à un facteur (« genre du locuteur ») a été conduite sur les données relatives aux locuteurs francophones. Ce test fait état d'un effet significatif de ce facteur sur la durée moyenne des mots : $F_{(1,268)}$=48,94 ; p<0,0001. Une analyse similaire conduite sur les mots produits par les anglophones américains révèle également un effet significatif du genre du locuteur : $F_{(1,268)}$=200,28 ; p<0,0001. Ces résultats confirment donc l'utilisation d'un débit de parole significativement plus élevé chez les locuteurs masculins dans les deux langues.

3.3 Répartition temporelle C/V

La répartition consonnes / voyelles au sein des mots CVCV produits par les locutrices et locuteurs francophones est présentée dans la Table 6, ci-dessous.

	Proportion (%)	
	Consonnes	**Voyelles**
Femmes	53,71	46,29
Hommes	45,93	54,07

TABLE 6 – Proportion moyenne (%) occupée par les consonnes et les voyelles au sein des 24 mots dissyllabiques de type CVCV pour les femmes et les hommes francophones, calculée sur 120 items (24 mots * 5 locuteurs) par genre.

On remarque ici une importante différence inter-genres : les consonnes sont proportionnellement plus longues chez les locutrices que chez les locuteurs francophones. Elles représentent environ 54% de la durée totale moyenne des mots chez les femmes, contre seulement 46% chez les hommes. Une ANOVA à un facteur (« genre du locuteur ») sur la proportion de chaque mot dissyllabique occupée par les consonnes conclue à une forte significativité de la différence inter-genres : $F_{(1,238)}$=92,699 ; p<0,0001.

La répartition C/V sur les mots dissyllabiques produits par les locuteurs anglophones américains est visible dans la Table 7, ci-après.

	Proportion (%)	
	Consonnes	**Voyelles**
Femmes	47,64	52,54
Hommes	45,45	54,32

TABLE 7 – Proportion moyenne (%) occupée par les consonnes et les voyelles au sein des 24 mots dissyllabiques de type CVCV pour les femmes et les hommes anglophones américains, calculée sur 120 items (24 mots * 5 locuteurs) par genre.

Chez les locuteurs anglophones, on observe une légère différence femmes-hommes sur la répartition temporelle entre consonnes et voyelles à l'intérieur des 24 mots de type CVCV : ici encore, les consonnes sont proportionnellement plus longues chez les femmes que chez les hommes. Une ANOVA à un facteur (« genre du locuteur ») confirme l'existence d'une différence inter-genre significative sur ce point : $F_{(1,228)}$=6,712 ; p=0.0102. Cependant, cette différence reste moins prononcée que chez les francophones.

4. Discussion - Conclusion

Concernant le VOT des consonnes occlusives, d'importantes différences inter-genres sont apparues. Chez les locuteurs francophones, le VOT des femmes est significativement plus long que celui des hommes sur les occlusives sourdes, et significativement plus court sur les occlusives voisées. Le contraste entre ces deux types de consonnes est donc très nettement accentué chez les locutrices par rapport aux locuteurs masculins. Pour les anglophones américains, le VOT est significativement plus long chez les femmes sur les occlusives aspirées, ce qui va dans le sens des observations faites notamment par Swartz (1992), Whiteside & Irving (1997) ou encore Robb et al. (2005). Sur les occlusives non-aspirées, même si cette différence n'est pas significative, le VOT des locutrices est également plus long que celui des locuteurs. Cependant, le contraste entre ces deux types de consonnes demeure, comme pour les francophones, significativement plus grand chez les locutrices. Ainsi, dans les deux langues, les femmes marquent une distinction plus forte que les hommes entre les deux types de consonnes occlusives (voisées / non-voisées ou aspirées / non-aspirées). Ces différences pourraient relever de facteurs culturels et socio-phonétiques, notamment d'une tendance à une articulation plus « soignée » chez les locutrices (Simpson, 2009).

Outre le VOT, plusieurs autres paramètres temporels ont été analysés, à commencer par la durée globale des mots dissyllabiques. Cette dernière est significativement plus élevée chez les locutrices que chez les locuteurs masculins. Cela rejoint des constatations faites dans plusieurs études antérieures (Byrd, 1994 ; Whiteside, 1995; Fitzsimons et al., 2001). L'ampleur de cette différence inter-genres est ici nettement plus grande chez les anglophones que chez les francophones. Néanmoins, ces résultats suggèrent que l'utilisation d'un débit de parole plus élevé chez les locuteurs masculins pourrait être une caractéristique assez largement partagée à travers les langues. Une différence inter-genres sur ce paramètre ne peut avoir qu'une origine sociologique et/ou culturelle : elle pourrait relever, là encore, d'une tendance à une articulation plus minutieuse chez les locutrices, tout du moins lors d'une tâche expérimentale de lecture, comme c'était le cas dans la présente étude. Une autre piste d'explication doit également être considérée : les locuteurs masculins tendraient à parler avec un débit rapide et en réalisant peu de pauses afin de dominer les conversations, en réduisant ainsi les possibilités d'interruption et de prises de parole par les autres interlocuteurs (Whiteside, 1995).

Autre élément intéressant, on constate que les consonnes occupent une proportion temporelle plus importante du mot chez les femmes, tant pour les francophones (52 % contre 46 %) que pour les anglophones (48 % contre 43 %). Les locutrices tendraient donc à faire durer plus longtemps les consonnes des mots (et donc moins longtemps les voyelles) que ne le font leurs homologues masculins. Ces différences, qui contredisent en partie une étude menée précédemment par Simpson et Ericsdotter (2003) sur des anglophones américains, relèvent très probablement d'habitudes articulatoires d'ordre socio-phonétique. Ces données pourraient être liées, une nouvelle fois, à la recherche d'une plus grande d'intelligibilité chez les locutrices : Owren et Cardillo (2006) ont en effet mis en évidence le rôle primordial joué par les consonnes dans l'identification des mots par l'auditeur.

Néanmoins, ces résultats doivent être interprétés avec prudence. Malgré des critères de sélection des participants extrêmement précis et la très faible variation intra-genre observée dans les données recueillies, la taille de l'échantillon demeure faible (10 locuteurs par langue). Il conviendra donc à l'avenir de poursuivre cette étude avec un nombre plus élevé de participants. De plus, on sait que le type de tâche effectué par les locuteurs influence fortement certains paramètres acoustiques, tels que le débit de parole ou la qualité de voix. Le corpus était ici composé de mots dissyllabiques lus : il conviendra de vérifier si des résultats similaires sont obtenus avec du discours spontané ou semi-spontané.

Références

BYRD, D. (1994). Relations of sex and dialect to reduction. *Speech Communication*, 15, 39-54.

BOË, L.-J., CONTINI, M., & RAKOTOFIRINGA, H. (1975). "Étude statistique de la fréquence laryngienne", *Phonetica*, 32: 1-23.

DI CRISTO, A. (1999). Vers une modélisation de l'accentuation du français : première partie. *Journal of French Language Studies*, 9, 143–179.

FITZSIMONS, M., SHEAHAN, N. & STAUNTON, H. (2001). Gender and the integration of acoustic dimensions of prosody: implications for clinical studies. *Brain and Language*, 78, 94-108.

FOX, R. A. & NISSEN, S. L. (2005). Sex-related acoustic changes in voiceless English fricatives. *Journal of Speech, Language, and Hearing Research*, 48, 753–765.

JACEWICZ, E., FOX, R. A., O'NEILL, C. & SALMONS, J. (2009). Articulation rate across dialect, age, and gender. *Language Variation and Change*, 21, 233-256.

JOHNSON, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, 34, 485‑499.

KARLSSON, F., ZETTERHOLM, E. & SULLIVAN, K. (2004). Development of a gender difference in voice onset time. In *Proceedings of the 10th Australian International Conference on Speech Science & Technology*, Sydney, 316-321.

KAHANE, J. (1978). A morphological study of the human prepubertal and pubertal larynx. *American Journal of Anatomy*, 151, 11-20.

LIN, C. & WANG, H. (2011). Automatic estimation of voice onset time for word-initial stops by applying random forest to onset detection, *Journal of the Acoustical Society of America*, 130, 514-525.

OH, E. (2011). Effects of speaker gender on voice onset time in Korean stops. *Journal of Phonetics*, 39, 59-67.

OWREN, M. J. & CARDILLO, G. C. (2006). The relative roles of vowels and consonants in discriminating talker identity versus word meaning. *Journal of the Acoustical Society of America*, 119, 1727-1739.

ROBB, M., GILBERT, H. & LERMAN, J. (2005). Influence of gender and environmental setting on voice onset time. *Folia Phoniatrica et Logopaedica*, 57, 125-133.

SIMPSON, A. P. (2009). Phonetic differences between male and female speech. *Language and Linguistics Compass*, 3, 621-640.

SIMPSON, A. P. & ERICSDOTTER, C. (2003). Sex-specific durational differences in English and Swedish. In *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona*, 1113-1116.

SWARTZ, B. L. (1992). Gender difference in voice onset time. *Perceptual and Motor Skills*, 75, 983-992.

WHITESIDE, S. P. (1995). Temporal-based speaker sex differences in read speech: A sociophonetic approach. In *Proceedings of the 13th International Congress of Phonetic Sciences*, Stockholm, 516-519.

WHITESIDE, S. P. (2001). Sex-specific fundamental and formant frequency patterns in a cross-sectional study. *Journal of the Acoustic Society of America*, 110, 464-478.

WHITESIDE, S. P. & IRVING, C. J. (1997). Speakers' sex differences in voice onset time, some preliminary findings. *Perceptual and Motor Skills*, 85, 459-463.

Voyelles moyennes en français calédonien : propriétés phonétiques acoustiques

Eleanor Lewis
Laboratoire de Phonétique, l'Université de Melbourne, Victoria 3010, Australie
elewis@unimelb.edu.au

RESUME

Cette étude examine la réalisation des voyelles moyennes /e, ɛ, ø, œ, o, ɔ/ par dix locuteurs du français calédonien. Les propriétés formantiques de ces voyelles sont analysées en ce qui concerne le genre de syllabe dans lesquelles elles se produisent. La durée des voyelles mi-fermées et mi-ouvertes produites en paires minimales est statistiquement comparée. Les résultats indiquent que les locuteurs de cette variété ont tendance à respecter catégoriquement la loi de position, tel que les variantes mi-fermées se présentent dans les syllabes ouvertes et les variantes mi-ouvertes se présentent dans les syllabes fermées. Il existe pourtant une certaine variation individuelle concernant le niveau de conformité à cette loi. Cette étude met également en avant des indices de l'antériorisation du /ɔ/ (et du /o/ en syllabe fermée), une caractéristique qui a été documentée dans d'autres variétés du français.

ABSTRACT

Mid vowels in New Caledonian French: Acoustic phonetic properties
This study examines the realisation of the mid vowels /e, ɛ, ø, œ, o, ɔ/ by ten speakers of New Caledonian French. The formant properties of these vowels are analysed with respect to the type of syllable in which they occur, and the durations of close-mid and open-mid vowels produced in minimal pairs are statistically compared. Results indicate that speakers of this variety tend to comply categorically with the *loi de position*, such that close-mid variants occur in open syllables and open-mid variants occur in closed syllables. There is some interspeaker variation, however, regarding the level of compliance with this rule. This study also highlights some evidence of fronting of /ɔ/ (and /o/ in closed syllables), a feature that has been documented in other varieties of French.

MOTS-CLES : Nouvelle-Calédonie, français calédonien, variation régionale, phonétique acoustique, voyelles moyennes
KEYWORDS: New Caledonian French, regional variation, acoustic phonetics, mid vowels

1 Introduction

La Nouvelle-Calédonie est une collectivité *sui generis* de France comptant environ 269 000 habitants[1], située dans le sud de l'océan Pacifique, à environ 1500 km de la côte est de l'Australie, mais à près de 17 000 km de la France métropolitaine. Le français calédonien (FC) est une variété régionale du français qui est peu documentée dans la linguistique, particulièrement en ce qui

[1] Recensement de 2014, voir http://www.isee.nc/population/recensement

concerne sa phonétique et sa phonologie. Les quelques descriptions existantes du système vocalique du FC (par ex. Hollyman, 1964, 1979, Pauleau, 1988, 2013) ont souligné une différence entre cette variété et le français standard au niveau des voyelles moyennes, /e, ɛ, ø, œ, o, ɔ/.

1.1 Les voyelles moyennes françaises

Le français standard est généralement considéré comme disposant de deux séries phonémiques de voyelles d'aperture moyenne : les mi-fermées /e, ø, o/ et les mi-ouvertes /ɛ, œ, ɔ/ (Fagyal et al., 2006, Fougeron, Smith, 1993). La réalisation de ces voyelles est compliquée par la « loi de position » (LdP), une règle de distribution complémentaire selon laquelle les variantes mi-ouvertes se produisent dans les syllabes fermées et les variantes mi-fermées apparaissent dans les syllabes ouvertes (Fagyal et al., 2006, Gess et al., 2012). Cette loi touche seulement les voyelles dans les syllabes finales des mots. En effet, celles dans les syllabes non-finales sont réalisées avec une qualité intermédiaire ou bien sont affectées par l'harmonie vocalique (voir par exemple Fagyal, Nguyen, et al., 2002). La LdP est plus une tendance qu'une loi stricte dans la plupart des variétés du français métropolitain. Des exceptions se présentent sous la forme de paires minimales comprenant les oppositions /ø/-/œ/ et /o/-/ɔ/ dans les syllabes fermées (par ex. *jeûne* /ʒøn/ vs. *jeune* /ʒœn/, *saute* /sot/ vs. *sotte* /sɔt/) et l'opposition /e/-/ɛ/ dans les syllabes ouvertes (par ex. *des* /de/ vs. *dais* /dɛ/). Dans d'autres variétés du français, notamment celles parlées dans le sud de la France, la LdP est plus strictement suivie et ces paires minimales deviennent homophones (c.-à-d. *saute* et *sotte* tous deux prononcés en [sɔt], *des* et *dais* tous prononcés en [de]) (Coquillon, Turcson, 2012, Durand, 2009, entre autres).

Il existe des indices du changement que les voyelles moyennes subissent actuellement dans les variétés métropolitaines, y compris dans le français parisien, vers une éventuelle perte des contrastes mi-fermés/mi-ouverts (Hansen, Juillard, 2011, Landick, 1995). Cette tendance s'observe particulièrement pour le contraste /e/-/ɛ/ (voir Fagyal, Hassa, et al., 2002 pour une étude acoustique de ce phénomène), tandis que le contraste /o/-/ɔ/ est le moins affecté par la neutralisation. Une croissance importante dans l'emploi de variantes intermédiaires (c.-à-d. des voyelles « moyennes » dans le sens plus étroit du terme) a également été documentée (Hansen, Juillard, 2011). Un deuxième phénomène touchant une de ces voyelles en français métropolitain est l'antériorisation de la voyelle postérieure /ɔ/ (à tel point qu'elle est rendue [ɔ̈] ou [œ]) (Armstrong, Low, 2008, Boula de Mareüil et al., 2013, Mooney, 2016). Cette antériorisation est une caractéristique longtemps attestée du français parisien de la classe ouvrière, qui s'est assez récemment répandue dans les variétés métropolitaines, atteignant même un certain prestige (Armstrong, Low, 2008).

1.2 Les voyelles moyennes en FC

Hollyman (1964) a émis l'hypothèse qu'il n'y a en français calédonien qu'une seule série phonémique de voyelles d'aperture moyenne (il utilise pour ces voyelles la notation /e, œ, o/), et que l'opposition mi-fermée/mi-ouverte est « remplacée par des variantes combinatoires ou positionnelles pour chaque voyelle moyenne » (Hollyman, 1979, p.623). Le mot « positionnelle » dans cette description peut faire référence à la variation allophonique selon la loi de position, pourtant cette hypothèse reste à étudier empiriquement. Pauleau (1988, 2013) ne s'est pas explicitement occupée du statut phonémique des voyelles moyennes en FC, mais elle a aussi constaté que ces voyelles diffèrent en aperture des mêmes voyelles en français métropolitain (non-méridional). Plus précisément, elle a noté une fermeture du /ɛ/ vers [e], une ouverture du /ø/ en [œ] et une perte du contraste /o/-/ɔ/ ou même l'inversion de leurs apertures.

1.3 Objectifs

À la vue des hypothèses offertes par les descriptions précédentes du FC et du comportement des voyelles moyennes françaises en général, cette étude vise à examiner la réalisation de ces voyelles en FC, traitant en particulier les questions suivantes :

1. Les locuteurs calédoniens : dans quelle mesure suivent-t-ils la LdP dans leur production des voyelles moyennes ?
2. Est-ce que les contrastes mi-fermés/mi-ouverts sont conservés dans les paires minimales, soit en qualité formantique, soit en durée ?

2 Méthode

2.1 Collecte de données

La parole de dix locuteurs calédoniens (7 femmes et 3 hommes), tous étudiants de premier cycle à l'Université de la Nouvelle-Calédonie âgés de 18 à 21 ans, a été enregistrée dans le cadre de cette étude. Tous les locuteurs ont grandi et ont effectué l'ensemble de leur scolarité en Nouvelle-Calédonie. Malgré le nombre assez restreint de locuteurs, le niveau de diversité ethnique et linguistique est considérable, comme on pourrait s'y attendre étant donnée la démographie de la Nouvelle-Calédonie. La plupart des participants sont des locuteurs natifs du français, cependant deux d'entre eux ont une langue maternelle kanak (nengone, kwenyii). Ces personnes sont bilingues en français, ayant suivi toutes leurs études en français. Quatre locuteurs de plus sont bilingues, soit en langue kanak (drehu, paicî), soit en langue asiatique (cantonnais, javanais). Les séances d'enregistrement ont eu lieu dans une salle de réunion relativement calme à l'Université de la Nouvelle-Calédonie. Un microphone tour d'oreille AudioTechnica AT892c et un enregistreur portable H4n Zoom ont été utilisés pour les enregistrements, avec une fréquence d'échantillonnage à 44,1 kHz et une résolution de 16 bits.

Les locuteurs ont produit une combinaison de mots réels et inventés comprenant toutes les voyelles du français standard dans le contexte /pVp/ (ou /pV/ pour les voyelles qui n'apparaissent pas dans les syllabes fermées) et des vrais mots additionnels contenant chaque voyelle. Des paires minimales supplémentaires pour les voyelles mi-fermées et mi-ouvertes /e/-/ɛ/, /ø/-/œ/ et /o/-/ɔ/ ont également été ajoutées. Les mots choisis ont été placés dans une phrase cadre et ont été répétés isolés à la fin de la phrase (« Je dis X encore. X » pour les mots de la forme /CVC/ ou /CVCC/ et « Je dis X parfois. X » pour ceux de la forme /CV/). Les phrases ont été présentées aux locuteurs sous la forme d'une présentation PowerPoint où chaque phrase (suivi par le même mot cible isolé) a occupé une diapositive. Les phrases sont apparues dans un ordre aléatoire et chaque phrase a été répétée 4 fois au cours au cours de la tâche. Tous les 35 mots ont été prononcés donc 8 fois (4 fois entourés par la phrase cadre et 4 fois isolés) et 2777 voyelles au totale ont été incluses dans l'analyse (23 occurrences ont été rejetées pour des raisons techniques).

2.2 Analyse acoustique

Les voyelles cibles ont été segmentées et étiquetées manuellement sous *Praat* (Boersma, Weenink, 2011), à l'aide de repères acoustiques pertinents. En plus du début et de la fin des voyelles, pour chaque occurrence, l'« état stable », c'est-à-dire la région la plus forte de stabilité formantique, a été identifiée et marquée. Le codage des segments a été fait selon les phonèmes attendus dans les mots cibles en français standard (ceux notés dans les dictionnaires ou ceux provoqués par l'orthographe).

Après ce traitement initial, les fichiers ont été convertis pour un traitement ultérieur dans le logiciel *Emu Speech Database System* (Harrington, 2010), qui permit la création de hiérarchies entre les niveaux d'étiquettes. À l'aide de ce logiciel, les traces des formants ont été examinées et rectifiées manuellement dans le cas d'erreurs.

En utilisant le paquet *Emu/R* dans le logiciel statistique *R* (Ihaka, Gentleman, 1996), les valeurs (en hertz) des deux premiers formants de toutes les voyelles ont été extraites aux points médians des états stables (ici appelées les « cibles vocaliques »). Ces cibles ont été placées dans un graphe F1~F2. Afin de quantifier cette information visuelle, des distances Euclidean et des comparaisons ERatio ont été subséquemment calculées (selon la procédure employée par Harrington, 2010). Ces mesures ont été utilisées pour chaque occurrence d'une voyelle susceptible à être touchée par la loi de position, en vue de déterminer si elle était plus proche dans l'espace F1~F2 au centroïde de la voyelle mi-fermée en syllabe ouverte (comme elle a été produite par le même locuteur) ou à celui de la voyelle mi-ouverte réalisée en syllabe fermée. Autrement dit, pour vérifier si la loi de position était respectée ou si les phonèmes ont été prononcés de la même façon dans les deux contextes syllabiques. Finalement, la durée des voyelles apparaissant dans les paires minimales a été calculée, et ces valeurs ont été comparées (utilisant un test-t apparié, suite au résultat d'un test de normalité Shapiro-Wilk) afin d'établir si les contrastes phonémiques ont été conservés grâce à une différence de durée.

3 Résultats

3.1 Cibles vocaliques

La Figure 1 (ci-dessous) montre les cibles vocaliques moyennes (F1~F2) produites par les dix locuteurs du FC (tous locuteurs confondus, graphes séparés pour les femmes et les hommes)[2]. Les phonèmes mi-fermés sont présentés en bleu foncé quand ils sont réalisés dans les syllabes ouvertes et en bleu clair quand ils sont dans les syllabes fermées. Les mi-ouverts sont en violet quand ils apparaissent en syllabe fermée et en rouge pour les syllabes ouvertes (ceci est seulement le cas pour le /ɛ/). Ce qui est frappant dans les deux graphes est la proximité des voyelles mi-fermées /ø/ et /o/ en syllabes fermées (violet) avec respectivement les voyelles équivalentes mi-ouvertes /œ/ et /ɔ/ réalisées dans le même contexte (bleu clair), par rapport aux mêmes phonèmes dans les syllabes ouvertes (bleu foncé). De la même manière, la voyelle mi-ouverte /ɛ/ en syllabe ouverte est beaucoup plus proche à la voyelle /e/ dans le même contexte syllabique (bleu foncé) qu'au même phonème comme il est produit dans les syllabes fermées (violet). Ces proximités démontrent que, pour ces locuteurs, les paires minimales différenciées par les contrastes mi-fermés/mi-ouverts deviennent homophones la plupart du temps (par ex. *des* et *dais* tous deux réalisés en [de], *saute* et *saute* en [sɔt], *jeûne* et *jeune* en [ʒœn]). Ce qui apparaît également évident dans les graphes, c'est la position antériorisée des voyelles /ɔ/ et /o/ dans les syllabes fermées (une position semblable ou même plus en avant à celle observée pour le /ɔ/ dans le français métropolitain du nord par Boula de Mareuïl et al., 2013, par exemple).

Les Figure 2 et 3 présentent quatre exemples de graphes de locuteurs individuels (aux ellipses d'intervalles de confiance de 95%) pour démontrer la distribution des voyelles dans l'espace F1~F2 (plutôt que la moyenne seulement). Bien qu'il existe des différences dans les spécificités phonétiques de certaines voyelles, pour la majorité des locuteurs, on observe un chevauchement

[2] Il faut rappeler que les locutrices sont 7 alors que les locuteurs ne sont que 3, c'est pourquoi les résultats pour les femmes sont plus fiables que ceux pour les hommes

entier entre les distributions des voyelles /ø/ et /œ/, et /o/ et /ɔ/, en syllabe fermée (ellipses bleu clair et violettes), et de /e/ et /ɛ/ en syllabe ouverte (ellipses bleu foncé et rouges). L'antériorisation du /ɔ/ et du /o/ en syllabe fermée notée dans les cibles vocaliques moyennes devient encore plus évidente dans les graphes individuels, qui montrent que certaines occurrences de ces voyelles sont réalisées avec les mêmes valeurs F1 et F2 que celles des voyelles /ø/ et /œ/.

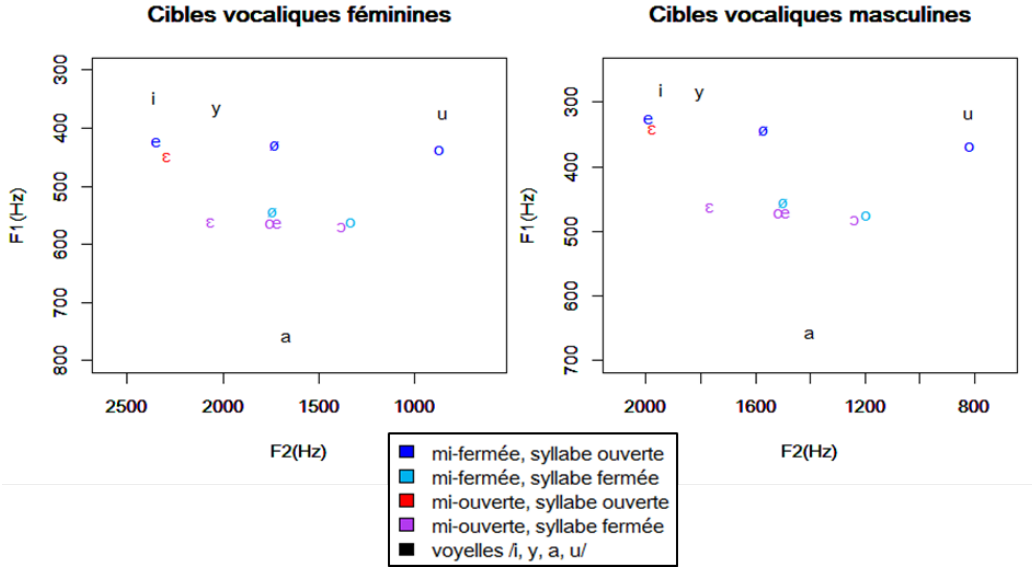


FIGURE 1 : Cibles vocaliques moyennes (F1~F2) des locuteurs féminins (à gauche) et masculins

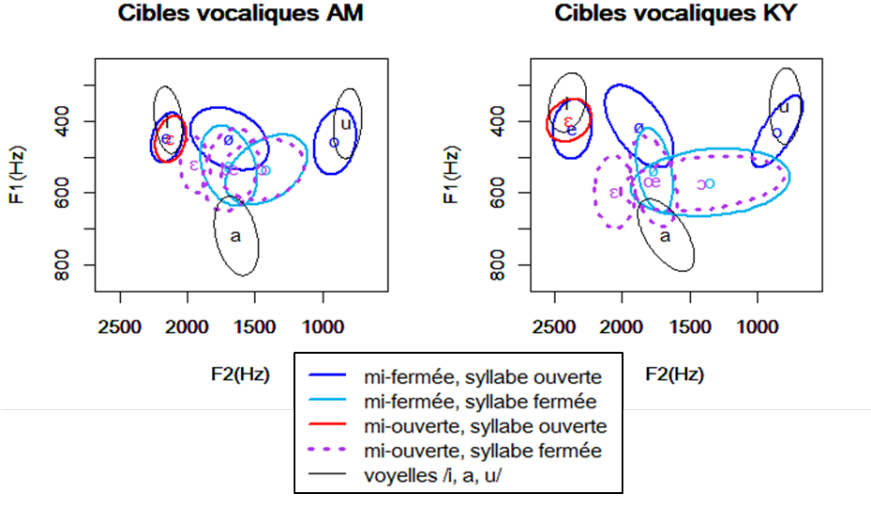


FIGURE 2 : Cibles vocaliques (F1~F2) (95% IC) des locutrices AM et KY

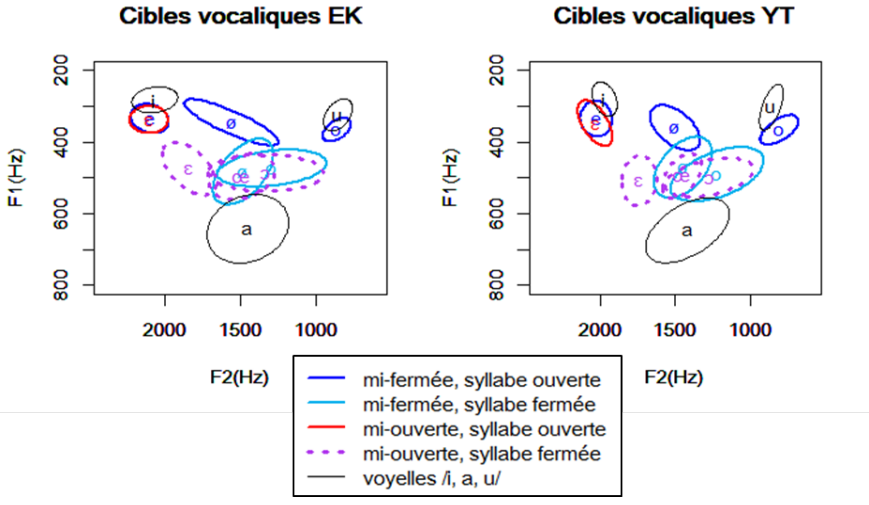


FIGURE 3 : Cibles vocaliques (F1~F2) (95% IC) des locuteurs EK et YT

3.2 Distances Euclidean et ERatios

Les tendances découvertes dans les graphes au-dessus sont soutenues par les distances Euclidean entre les voyelles pertinentes et les comparaisons ERatio de celles-ci. Les valeurs moyennes ERatio, séparées par voyelle et par locuteur, sont présentées sur la Table 1. Un chiffre positif indique que la voyelle est généralement réalisée d'une façon similaire dans les deux contextes syllabiques, alors qu'un chiffre négatif montre que la voyelle est plutôt similaire à la voyelle de l'autre aperture dans le même contexte (un chiffre plus grand indique que la tendance est plus forte). Un résultat autour de zéro suggère qu'une voyelle est réalisée avec une qualité intermédiaire.

Locuteur	ERatio moyen /ɛ/ syllabe ouverte	ERatio moyen /ø/ syllabe fermée	ERatio moyen /o/ syllabe fermée
AD (f)	-1.49	-0.86	-1.02
MA (f)	**1.5**	-1.1	-1.42
KY (f)	-1.9	-0.87	-0.96
JV (f)	-2.69	-1.04	-1.68
KS (f)	-1.46	**0.03**	-1.43
EV (f)	-0.69	-0.45	-0.53
AM (f)	-1.54	-0.38	-1.71
AG (m)	-0.67	-0.9	-1.33
EK (m)	-2.12	-0.86	-1.13
YT (m)	-2.22	-0.79	-1.57

TABLE 1 : ERatios moyens par voyelle et par locuteur

D'après ces résultats, la qualité des voyelles moyennes suit la loi de position la plupart du temps pour ces locuteurs. Cela se reflète dans les valeurs moyennes ERatio majoritairement négatives. Les seuls moyennes positives (marquées en gras sur la table) sont celles du /ɛ/ en syllabe ouverte pour la locutrice MA (1.5) et du /ø/ en syllabe fermée pour la locutrice KS (0.03). Cette valeur première considérablement positive montre que cette locutrice conserve le contraste /e/-/ɛ/ dans les syllabes ouvertes (c.-à-d. *des* [de] et *dais* [dɛ]). La deuxième valeur, par contraste, est un chiffre très faible, ce qui indique probablement une réalisation d'aperture intermédiaire pour la voyelle /ø/ en syllabe fermée pour la locutrice KS.

3.3 Durée

Les analyses de la durée des voyelles moyennes en paires minimales sont présentées sur la Table 2. Les test-t tous locuteurs confondus révèlent des différences statistiquement significatives entre celles de /ø/ et /œ/ ($p = 0.012$) et celles de /o/ et /ɔ/ ($p = 0.048$). Cependant, les écarts eux-mêmes sont très modestes (4.66 ms et 1.86 ms respectivement) et il semble donc peu probable qu'ils soient significatifs (dans le sens plus large du mot) pour les locuteurs et leurs auditeurs (cf. par exemple Lehiste, 1976 au sujet du seuil de discrimination pour la durée)[3]. Aucune différence significative n'est observée entre les durées des voyelles /e/ et /ɛ/.

Voyelles	**/e/-/ɛ/**	**/ø/-/œ/**	**/o/-/ɔ/**
Mi-fermée - durée moy. (ms)	150.99	139.64	107.47
Mi-ouverte - durée moy. (ms)	159.48	134.98	105.61
Ecart (ms)	8.49	4.66	1.86
Valeur-p	0.251 n.s.	0.012 *	0.048*

TABLE 2 : Comparaisons de durée des voyelles moyennes en paires minimales

4 Discussion et conclusion

Les résultats ici présentés démontrent un respect général de la loi de position dans la réalisation des voyelles moyennes en français calédonien (comme on le voit également dans le français du Midi, par exemple). Ainsi, les variantes mi-fermées apparaissent dans les syllabes ouvertes alors que les variantes mi-ouvertes apparaissent dans les syllabes fermées, tendance qui rend homophones les paires minimales distinguées par les oppositions mi-fermées/mi-ouvertes (par exemple *tes/tais* [te], *côte/cote* [kɔt], *veule/veulent* [vœl]). Cela est apparent dans les graphes des cibles vocaliques puisqu'ils montrent que les moyennes des voyelles /e/-/ɛ/, /ø/-/œ/ et /o/-/ɔ/ sont toutes à côté les unes des autres quand elles sont produites dans les mêmes contextes syllabiques. Ce fait est quantifié à travers les ERatio des distances Euclidean entre les voyelles pertinentes. La tendance à la suite de la LdP est claire aussi dans les comparaisons de la durée pour les voyelles moyennes réalisées dans les paires minimales, qui diffèrent de façon significative, mais tellement peu que ces différences seraient probablement imperceptibles aux auditeurs. Il semble peu plausible donc que les contrastes phonémiques soient conservés en employant une différence de durée. Ainsi, les résultats semblent soutenir la proposition de Hollyman (1964, 1979) selon laquelle il existe une seule série de voyelles moyennes en français calédonien (réalisée par des variantes « positionnelles »), ainsi que la

[3] Il serait néanmoins utile de vérifier cette supposition à travers une future expérience de perception

déclaration de Pauleau (1988, 2013) que l'aperture de ces voyelles diffère dans cette variété régionale.

Une deuxième découverte de cette étude est l'antériorisation fréquente des voyelles /ɔ/ et /o/ (en syllabe fermée), un phénomène qui a été déjà documenté dans plusieurs variétés du français métropolitain (Armstrong, Low, 2008, Boula de Mareüil et al., 2013, Mooney, 2016). Dans les variétés métropolitaines, l'antériorisation du /ɔ/ se fait plus fréquemment dans les contextes qui provoquent la coarticulation dans cette direction, par exemple les consonnes contigües apicales. Ici également l'antériorisation de ces voyelles peut être une conséquence de l'effet coarticulatoire d'un [t] suivant (dans les mots *côte/cote* et *saute/sotte*) (ou du [s] précédant dans le cas de *saute/sotte*).

Cette étude correspond à une étape préliminaire dans l'analyse des voyelles moyennes en français calédonien, et de ce fait elle a des limitations. Rappelons, par exemple, que les données ont été récoltées lors d'une tâche de lecture, et que l'éventail d'âge dans l'échantillon des locuteurs restait plutôt limité. Des futures études considéreront donc les voyelles moyennes produites dans des conditions d'expression moins formelles et par des locuteurs de plusieurs tranches d'âge. En outre, il reste à étudier de façon acoustique expérimentale plusieurs caractéristiques du français calédonien proposées par les chercheurs précédents, dont une perte fréquente de l'opposition des voyelles nasales /ɑ̃/ et /ɔ̃/ (par ex. Hollyman, 1964, Pauleau, 1988). Contrairement au phénomène touchant les voyelles moyennes ici présenté, ce changement dans les voyelles nasales n'est pas répandu dans les variétés du français parlées en métropole ou ailleurs. Une prochaine étude acoustique examinera cette particularité du FC.

Remerciements

Je tiens à remercier vivement Sarah Anne Bégo pour son soutien linguistique dans l'écriture de cet article, ainsi que les trois relecteurs anonymes pour leurs commentaires utiles.

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.

BOERSMA, P., WEENINK, D. (2011). Praat: Doing phonetics by computer (Version 5.3) [Logiciel]. *http://www.praat.org*.

BOULA DE MAREÜIL, P., WOEHRLING, C., ADDA-DECKER, M. (2013). Contribution of automatic speech processing to the study of Northern/Southern French. *Language Sciences*, *39*, 75-82.

COQUILLON, A., TURCSON, G. (2012). An overview of the phonological and phonetic properties of Southern French. Dans R. Gess, C. Lyche & T. Meisenburg (éd.), *Phonological variation in French: Illustrations from three continents*. Philadelphia: John Benjamins, 105-127.

DURAND, J. (2009). Essai de panorama phonologique: les accents du Midi. Dans L. Baronian & F. Martineau (éd.), *Le français, d'un continent à l'autre. Mélanges offerts à Yves Charles Morin*. Québéc: Les presses de l'Université Laval, 123-170.

FAGYAL, Z., HASSA, S., NGOM, F. (2002). *L'opposition [e]-[ɛ] en syllabes ouvertes de fin de mot en français parisien: Etude acoustique préliminaire.* Actes des 24èmes Journées d'Etudes sur la Parole, Nancy, 165-168.

FAGYAL, Z., KIBBEE, D., JENKINS, F. (2006). *French: A linguistic introduction*. Cambridge: Cambridge University Press.

FAGYAL, Z., NGUYEN, N., BOULA DE MAREÜIL, P. (2002). From dilation to coarticulation: Is there vowel harmony in French? *Studies in the Linguistic Sciences, 32*(2), 1-21.

FOUGERON, C., SMITH, C. L. (1993). Illustrations of the IPA: French. *Journal of the International Phonetic Association, 23*(2), 73-76.

GESS, R., LYCHE, C., MEISENBURG, T. (2012). Introduction to phonological variation in French. Dans R. Gess, C. Lyche & T. Meisenburg (éd.), *Phonological variation in French: Illustrations from three continents*. Philadelphia: John Benjamins, 1-19.

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*(3), 313-359.

HARRINGTON, J. (2010). *Phonetic analysis of speech corpora*. Chichester: Wiley-Blackwell.

HOLLYMAN, K. J. (1964). *Le français régional de l'Indo-Pacifique: Essais de phonologie*. Auckland: Linguistic Society of New Zealand.

HOLLYMAN, K. J. (1979). Le français en Nouvelle-Calédonie. Dans A. Valdmann (éd.), *Le français hors de France*. Paris: Champion, 621-629.

IHAKA, R., GENTLEMAN, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5*(3), 299-314.

LANDICK, M. (1995). The mid-vowels in figures: Hard facts. *The French Review, 69*(1), 88-102.

LEHISTE, I. (1976). Suprasegmental features of speech. Dans N. J. Lass (éd.), *Contemporary issues in experimental phonetics*. New York: Academic Press, 225–239.

MOONEY, D. (2016). 'C'est jeuli, la Gasceugne!': l'antériorisation du phonème /ɔ/dans le français régional du Béarn. *French Studies, 70*(1), 61-81.

PAULEAU, C. (1988). *Étude phonétique contrastive du français calédonien et du français standard.* (Thèse de master), Université Paris 3.

PAULEAU, C. (2013). Description et sauvegarde du patrimoine immatériel de la langue franco-calédonienne. *Bulletin de la Société d'Etudes Historiques de la Nouvelle-Calédonie, 175*, 53-68.