Comparaison d'analyses phonétiques de parole dysarthrique basées sur un alignement manuel et un alignement automatique

Cécile Fougeron[1], Nicolas Audibert[2], Corinne Fredouille[2], Christine Meunier[3], Cédric Gendrot[1], Olavo Panseri[1]

[1] Lab. de Phonétique et Phonologie, UMR 7018 CNRS-Paris3/Sorbonne Nouvelle, Paris, France
[2] Université d'Avignon, CERI/LIA, Avignon, France
[3] Laboratoire Parole et Langage, CNRS, Université Aix-Marseille, France
cecile.fougeron@univ-paris3.fr, corinne.fredouille@univ-avignon.fr, christine.meunier@lpl-aix.fr

ABSTRACT

The reliability of an automatic speech alignment procedure for the phonetic description of dysarthric speech is assessed through the comparison of durational and spectral measurements obtained from an automatic and a manual alignment of the production of 4 dysarthric speakers varying in severity. Results show that formant values computed in the middle of the vowel intervals and center of gravity of fricative noise computed over the consonant intervals, are reliable when based on automatic alignments. However, the analysis of pause occurrences and absolute segmental duration require manual corrections of the automatic outputs.

Keywords: dysarthric speech, phonetico-acoustic study, automatic vs. manual alignment.

1. INTRODUCTION

La dysarthrie est un terme générique définissant un trouble de la parole d'origine motrice consécutif à une atteinte du système nerveux central et/ou périphérique. En fonction de la localisation de l'atteinte dans le cerveau, de la sévérité de la maladie associée, ou de particularités propres au locuteur, les caractéristiques de la dysarthrie varient. Ainsi, si toutes les dimensions de la parole (phonation, articulation, timing, prosodie, fluence…) peuvent être altérées, elles le sont à des degrés variables en fonction des patients. De plus, en fonction de la nature du trouble moteur, les altérations sur ces dimensions n'ont pas la même forme : un mauvais contrôle temporel des mouvements affectera la durée et les transitions entre segments, alors qu'une perte de force musculaire affectera plutôt l'amplitude de mouvement et l'atteinte des cibles articulatoires. Au vu de cette variabilité des formes de dysarthrie, nos recherches ont pour but de décrire finement les profils dysarthriques sur la base de leurs caractéristiques phonético-acoustiques, en isolant des critères robustes, quantifiables et surtout plus objectifs que des critères perceptifs dont se servent les classifications antérieures [1].

Les travaux présentés ici visent à établir une procédure d'analyse optimale en termes de temps et d'expertise humaine permettant l'analyse phonético-acoustique d'un grand nombre d'échantillons de parole dysarthrique pour faire face à la variabilité inter- et intra-locuteurs importante dans ce type de populations. La segmentation manuelle d'un continuum de parole est extrêmement coûteuse en temps et en expertise, et les altérations phonétiques dans la dysarthrie rendent ce travail encore plus ardu. Le recours à une segmentation automatique des productions apparaît donc comme une alternative des plus intéressantes. Pour autant, on sait que les systèmes d'alignement automatique peuvent générer des erreurs dans la localisation des frontières de phonèmes sur de la parole normale. Il est donc nécessaire d'une part, d'évaluer ces erreurs par rapport à une segmentation manuelle (voir Audibert et al. [2]), et d'autre part d'évaluer l'adéquation d'une telle approche en regard de la validité des analyses phonético-acoustiques qu'elle permet dans le contexte spécifique de la parole dysarthrique.

Notre objectif ici est donc d'évaluer la validité d'une analyse phonétique basée sur un alignement automatique de productions dysarthriques, en la comparant à une analyse basée sur un alignement manuel de référence. Il s'agira de savoir si les analyses phonétiques faites à partir de l'alignement automatique sont suffisamment proches de celles effectuées à partir d'un alignement manuel pour décrire les mêmes tendances. En fonction du type de critère phonétique étudié, nous chercherons à déterminer si l'on peut reposer l'analyse sur un alignement tout-automatique ou si celui-ci nécessite une phase préalable de vérification et correction manuelle.

2. METHODE

2.1. Corpus et locuteurs

Le corpus est constitué d'une partie du texte 'Tic Tac' de la batterie de C. Chevrie-Müller lu par quatre patients dysarthiques atteints de maladies rares de surcharges lysosomales. Ces enregistrements font partie d'une étude en collaboration avec F. Sédel et N. Lévêque (Hôpital de la Pitié Salpêtrière). Les locuteurs, 2 hommes et 2 femmes, présentent des dysarthries de type mixte à des degrés de sévérité différents. Les patients M1A, F1C (avec M pour homme et F pour femme) sont légèrement dysarthriques ; les patients M2V, F2S ont des dysarthries sévères, marquées plus particulièrement par des altérations de la qualité vocale

et un débit articulatoire ralenti chez M2V, et des altérations d'articulation consonantique et vocalique chez F2S.

2.2. Alignements automatique et manuel

Dans ce travail, le système automatique d'alignement contraint par le texte, développé par le Laboratoire Informatique d'Avignon (LIA), est utilisé. Ce système repose sur l'utilisation classique de modèles de Markov Cachés (38 modèles HMM indépendants du contexte estimés sur le corpus d'émissions radiophoniques ESTER, [3]), associés à un algorithme de décodage de type Viterbi [4]; une description plus détaillée du système est donnée dans Audibert *et al.* [2].

Les échantillons de parole étudiés ont été retranscrits manuellement sous une forme orthographique de façon à inclure toutes les insertions, suppressions, substitutions et répétitions produites par les patients par rapport au texte d'origine. Le lexique phonétisé, utilisé en entrée du système automatique, a été restreint aux seules entrées lexicales du texte lu, puis adapté dynamiquement à chaque transcription orthographique pour prendre d'éventuelles entrées manquantes (dues à des substitutions ou faux départs par exemple). Finalement, les variantes de prononciation de chaque entrée lexicale ont été vérifiées de façon à n'inclure que celles possibles dans le texte. Sur la base de la transcription orthographique fournie et du lexique phonétisé, le système automatique va analyser le signal de parole et identifier les frontières phonémiques (sous la forme d'étiquettes de début et de fin) de la séquence de phonèmes attendue.

Un alignement manuel (AM) a ensuite été réalisé par un expert humain sur la base de l'alignement automatique. Sa tâche consistait, d'une part, à vérifier la réalisation des phonèmes transcrits (donc éventuellement à changer, insérer ou supprimer des étiquettes de phonèmes) et, d'autre part, à déplacer les étiquettes de début et de fin lorsqu'il le jugeait nécessaire par rapport au signal produit. Le placement des frontières a été fait sur la base de critères de segmentation couramment utilisés : l'apparition et la disparition du 2^{e} formant pour les voyelles, le bruit caractéristique des fricatives, la tenue voisée ou silencieuse pour les plosives, le bruit correspondant au relâchement des occlusives, etc. Les difficultés majeures pour l'expert ont résidé dans le placement d'une frontière au sein de suites de voyelles et de consonnes de type vocalique ou sonant (comme dans "horloge", par exemple). Dans les cas les plus difficiles, l'expert a codé les portions de signal comme 'insegmentables'.

2.3. Procédure de comparaison

Afin de comparer les procédures manuelle et automatique sur les mêmes portions de signal, seuls les phonèmes segmentés dans les deux alignements, avec une étiquette phonémique similaire, ont été conservés. Les segments insérés ou supprimés dans l'un ou l'autre des alignements ont donc été exclus, ainsi que les parties du signal que l'expert humain a jugé comme 'insegmentables'. Les segments retenus ont été regroupés en grandes classes phonétiques par locuteur, et les classes présentant moins de 10 exemplaires par locuteur ont été éliminées (consonnes et voyelles nasales, semi-voyelles). Au final, 901 segments ont été retenus et leur distribution par classe phonétique est indiquée dans le tableau 1.

Tableau I : Nombre d'occurrences et distribution des segments comparés par locuteur et classe phonétique. F : fricative, O : occlusive, (b) : burst, (t) : tenue, Vo : voyelle orale, S : sourde, V : voisée, # : pause

Loc	FS	FV	OSb	OSt	OV	/R/	/l/	Vo	#
M1A	11	15	24	18	14	13	15	67	10
F1C	12	17	23	22	15	17	18	85	25
F2S	14	18	19	19	18	20	16	89	30
M2V	12	17	26	25	16	18	15	79	29

2.4. Critères phonétiques étudiés/comparés

Afin d'évaluer si l'utilisation d'un alignement automatique est adaptée pour l'étude de propriétés phonetico-acoustiques de la parole dysarthrique, différentes mesures acoustiques obtenues à partir de l'alignement manuel (AM) servent de référence. Elles sont comparées aux mesures obtenues à partir de l'alignement automatique (AA). Pour chaque mesure, l'effet du type d'alignement (AA vs. AM), et les interactions avec les facteurs 'locuteur' et 'classe phonétique' sont testés à l'aide d'ANOVAs.

Les comparaisons se basent sur quatre types d'analyses : mesures de la durée et du nombre de pauses, des durées segmentales, des fréquences des formants F1 et F2 des voyelles, du centre de gravité spectral (CoG) du bruit des fricatives. Le choix de ces mesures répond aux critères suivants. Premièrement, ces mesures peuvent caractériser différents aspects de la parole pouvant être altérés dans la dysarthrie : la fluence (pauses), la prosodie (durée et pauses), l'articulation des voyelles (durée, formants) et des consonnes (durée, CoG pour les fricatives). Ces mesures ont ainsi pu être utilisées dans la littérature dans le cadre de la description de parole pathologique (voir les revues dans [5] et [6]). Deuxièmement, elles touchent à des dimensions acoustiques différentes : spectrales pour les formants et le CoG, temporelles pour les durées. Un décalage temporel entre les étiquettes de l'AA et celles de l'AM peut avoir des répercussions sur des mesures de durée absolue des segments mais permettre la comparaison entre phonèmes ou entre locuteurs si les décalages sont systématiques. Les deux mesures spectrales sont calculées sur des empans différents : une mesure locale pour les formants pris au centre des voyelles, une mesure globale pour le CoG mesurée sur la fenêtre temporelle des fricatives.

3. RESULTATS

3.1. Nombre et durée des pauses

La comparaison de la durée des 94 intervalles segmentés comme des pauses (silences) par l'AA et l'AM ne montre pas d'effet du type d'alignement ($F(1,180)=1.59$, $p=.2$) : les durées issues des deux alignements sont similaires et ceci pour tous les locuteurs (interaction $F(3,180)=.82$, $p=.48$).

Pour autant, l'adéquation de l'alignement automatique pour l'analyse des pauses est illusoire. En effet, si toutes les pauses relevées par l'AM ont été détectées par l'AA, l'inverse n'est pas vrai. L'AA insère des pauses qui ne sont pas notées par l'expert humain et leur proportion n'est pas négligeable (64 insertions erronées sur le total des 4 locuteurs, pour 94 pauses réelles). Ces pauses apparaissent généralement dès lors que l'AA rencontre des difficultés pour aligner la séquence de phonèmes attendue sur le signal (soit parce que les phonèmes sont très dégradés, soit parce qu'ils sont trop longs et, par conséquent, qu'ils ne correspondent plus aux modèles de phonèmes du système). La présence de pauses optionnelles après chaque mot dans le lexique phonétisé offre la possibilité au système d'insérer une pause pour résoudre les incohérences temporelles qu'il rencontre. L'avantage de ce procédé est d'éviter de répercuter des erreurs d'alignement au-delà du mot. Néanmoins, ce dernier se fait au prix d'insertions erronées de multiples pauses dans l'alignement qu'il est donc nécessaire de corriger.

3.2. Durées segmentales

La comparaison effectuée sur les durées de 901 phonèmes extraites à partir des segmentations de l'AA et de l'AM montre un effet significatif du type d'alignement ($F(1,1550)=94.2$; $p<.0001$) : les durées de l'AA sont globalement plus courtes. Pour autant, cet effet varie en fonction du locuteur (interaction aligneur*locuteur $F(3, 1550)=20.15$, $p<.0001$) et de la classe de phonèmes (interaction aligneur*classe $F(7,1550)=18.28$, $p<.0001$).

En ce qui concerne les locuteurs, un effet du type d'alignement est trouvé chez les deux patients ayant une dysarthrie sévère (F2S et M2V) mais aussi chez la patiente à dysarthrie légère (F1C). Chez ces trois locuteurs, il y a une interaction avec la classe de phonèmes. Chez le locuteur H1A, les sorties de l'AA sont comparables à celles de l'AM.

Les résultats des analyses concernant les différentes classes de phonèmes sont illustrés sur la figure 1. Pour l'ensemble des fricatives l'AA donne des durées significativement plus courtes. Mais cet effet varie en fonction du locuteur. Comme illustré sur la figure 1, l'effet est notable pour les deux patients les plus dysarthriques (F2S et M2V) pour toutes les fricatives, et uniquement pour les fricatives voisées pour F1C. Pour les occlusives voisées et pour les voyelles, les durées de l'AA sont également significativement plus courtes et ceci pour tous les locuteurs. La tenue des occlusives est significativement plus longue dans l'AA avec une différence notable chez les deux patients les moins dysarthriques (H1A et F1C). L'explosion des occlusives est, quant à elle, plus courte dans l'AA chez tous les patients (mais avec un degré variable). Seules les durées des liquides /l/ et /ʁ/ ne diffèrent pas entre l'AA et l'AM. Ceci est particulièrement intéressant car ces segments sont souvent difficiles à délimiter sur le signal, même par des experts humains.

3.3. Formants des voyelles

Les occurrences de voyelles dans la partie du texte produite par nos patients ne sont pas nombreuses. Pour notre comparaison, nous avons élargi notre critère d'inclusion à au moins 9 exemplaires par type de voyelles pour chaque locuteur, de façon à pouvoir comparer 81 /a/, 43 /i/, 71 /e,ɛ/, 45 /o, ɔ/.

La comparaison des fréquences formantiques obtenues pour F1et F2 au centre de la voyelle ne montre pas d'effet du type d'alignement, ni d'interaction avec les facteurs 'locuteur' et 'type de voyelles'. Dans [2] nous montrons que le centre de la voyelle de l'AA et le centre de la voyelle de l'AM peuvent être décalés. Afin de déterminer quelles seraient les mesures optimales à appliquer sur la segmentation de l'AA, nous avons comparé les mesures formantiques prises au centre de l'AM (comme référence) à celles prises aux 1/3, 1/2 et 2/3 de l'intervalle vocalique segmenté par l'AA, et à une valeur moyenne calculée sur les 3 points. Pour chaque comparaison, une bonne corrélation est obtenue ($r=.9$ pour les mesures à 1/2, $r=.7$ pour 1/3 et $r=.7$ pour 2/3, $r=.9$ pour la moyenne), mais il semble toutefois plus prudent de prendre des mesures formantiques au centre de la voyelle ou une moyenne sur trois points avec un alignement automatique.

3.4. Centre de Gravité du bruit des fricatives

Nous avons comparé les mesures de CoG issues des deux alignements pour les fricatives en les regroupant selon des contrastes de lieu. On sait que le bruit des dentales est plus aigu que celui des fricatives labiales et post-alvéolaires. Afin d'avoir un nombre suffisant d'occurrences (au moins 10 par locuteur dans chaque catégorie), nous avons distingué articulation dentale ($N=92$) et articulation non-dentale ($N=140$). La comparaison des valeurs de CoG ne montre pas d'effet du type d'alignement. Les fricatives dentales ont un CoG plus haut pour tous les locuteurs et dans les deux alignements comme illustré sur la figure 2.

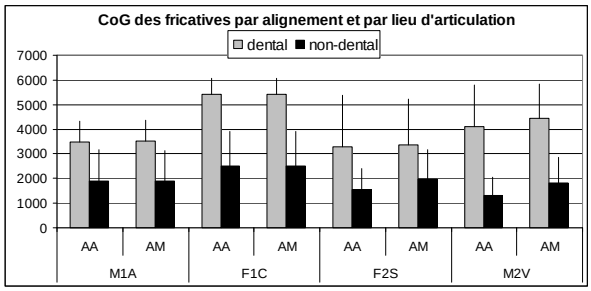

Figure 2 : Centre de gravité spectral (CoG en Hertz) des fricatives selon leur lieu d'articulation pour chaque patient et selon le type d'alignement.

4. DISCUSSION ET CONCLUSION

Il ressort de nos résultats que l'application d'une procédure d'alignement automatique pour l'analyse de critères phonético-acoustiques sur la parole dysarthrique est tout à fait envisageable, même si, en fonction des critères étudiés, elle nécessitera une vérification manuelle. Pour l'étude des pauses, nous avons vu que les segmentations automatiques doivent être vérifiées pour éliminer les insertions erronées de pauses. Pour l'étude de la durée des segments, la fiabilité d'un alignement automatique doit être appréciée en regard du type de segment à étudier et en fonction de la précision temporelle requise pour l'objet d'étude. Si l'on peut se reposer sur l'AA pour l'étude de la durée des liquides, une correction/vérification manuelle sera requise pour les autres segments. Par contre l'utilité d'une segmentation automatique n'est pas à exclure si l'étude des durées segmentales sert à examiner des contrastes entre types de phonèmes ou entre locuteurs. En effet, le contraste de durée entre fricatives sourdes et sonores apparaît aussi bien avec l'AM qu'avec l'AA (les fricatives sourdes sont plus longues que les sonores dans les deux alignements). De même, l'allongement des durées segmentales particulièrement important chez le patient M2V ressort dans les deux alignements. En ce qui concerne les mesures spectrales, les analyses basées sur une segmentation automatique semblent fiables aussi bien pour les formants des voyelles que pour les mesures de CoG sur le bruit des fricatives, segments dont la durée est sous-estimée par l'AA. Ces résultats sont particulièrement encourageants car ils montrent que ces analyses spectrales peuvent être effectuées directement sur l'AA, sans correction manuelle. Par ailleurs, la fiabilité de l'alignement automatique semble dépendre de la sévérité de la dysarthrie des locuteurs. En effet, les différences observées entre AA et AM sont plus fréquentes et importantes chez les patients les plus dysarthriques H2V et F2S. Toutefois, il faut noter que les experts humains sont également en difficulté face au signal de parole particulièrement altéré chez ce type de patients. Le nombre de séquences jugées insegmentables par les experts en témoigne.

Remerciements : *Ce travail est financé par l'ANR-08-BLAN-0125 et l'association Vaincre les Maladie Lysosomales. Nous remercions Georges Linares du LIA pour son aide sur le système d'alignement automatique.*

BIBLIOGRAPHIE

[1] F. L. Darley, A. E. Aronson and J. R. Brown. Clusters of Deviant Speech Dimensions in the Dysarthrias. *JSHR*, 12: 462-496, 1969.

[2] N. Audibert, C. Fougeron, C. Fredouille, C. Meunier and O. Panseri. Evaluation d'un alignement automatique sur la parole dysarthrique. In *Actes XXVIIIèmes JEP*, 2010.

[3] S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa and K. Choukri. Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. In *Proc. LREC'06*, 2006.

[4] F. Brugnara, D. Falavigna and M. Omologo. Automatic segmentation and labelig of speech based on hidden Markov models, *Speech Communication*, 12: 357-370, 1993.

[5] R. D. Kent, G. Weismer, J. F. Kent, H. K. Vorperian and J. R. Duffy. Acoustic studies of dysarthric speech: Methods, progress, and potential. *Journal of Com. Disorders*, 32/3:141-186, 1999.

[6] B. E. Murdoch. *Dysarthria*: *A physiological approach to assessment and treatment*. Stanley Thornes, Cheltenham, UK, 1998.

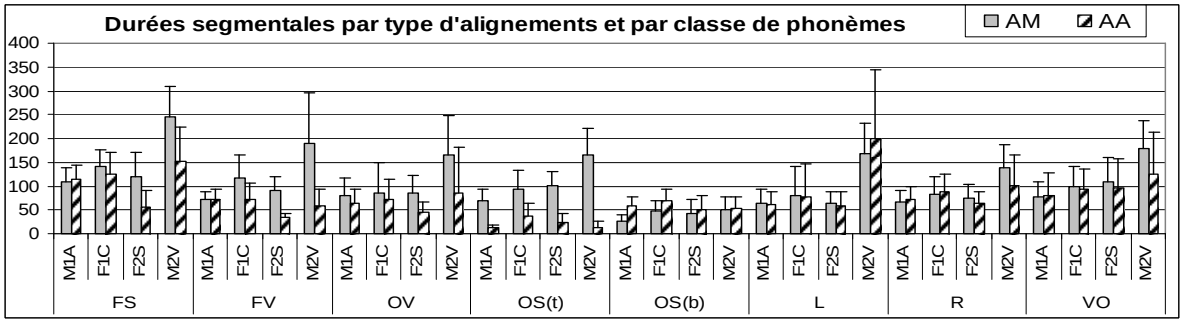

Figure 1 : durées (en ms) selon le type d'alignement pour chaque patient et pour chaque classe de phonèmes