Evaluation d'un alignement automatique sur la parole dysarthrique

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

[1] Université d'Avignon, CERI/LIA, Avignon, France
[2] Laboratoire de Phonétique et Phonologie, UMR 7018 CNRS-Paris3/Sorbonne Nouvelle, Paris, France
[3] Laboratoire Parole et Langage, CNRS, Aix-Marseille Université, France
{prenom.nom}@univ-avignon.fr, cecile.fougeron@univ-paris3.fr, christine.meunier@lpl-aix.fr, olavo.panseri@gmail.com

ABSTRACT

Phonetico-acoustic analysis of pathological speech requires a reliable phonetic alignment. Since manual labeling is highly time-consuming, automatic alignment may be necessary for analyzing large databases. This paper evaluates the reliability of automatic alignment for dysarthric speech. Results on read speech samples of 4 dysarthric speakers compared to 2 normophonic speakers show that alignment performance depends on the severity of dysarthria. Specific patterns for different phonetic classes and directions for filtering reliable parts are discussed.

Keywords: dysarthric speech, phonetico-acoustic study, automatic phonetic alignment, evaluation.

1. INTRODUCTION

Depuis une quinzaine d'années, la phonétique clinique, dédiée à l'étude des troubles de la parole, de la voix et du langage, est devenue un domaine multidisciplinaire où se côtoient cliniciens et chercheurs des domaines des sciences du langage et du traitement automatique de la parole.

La dysarthrie est un trouble de la parole d'origine neurologique qui se manifeste par une déficience motrice. Elle a fait l'objet de nombreuses études dans la littérature, portant notamment sur sa caractérisation dans le domaine acoustique ou donnant lieu à diverses classifications. Si un ensemble de paramètres jugés pertinents dans la différentiation de patients touchés par des types de dysarthries différents a pu être défini, la dysarthrie nécessite encore des études approfondies. Notamment, une description phonétique rigoureuse permettrait de mieux appréhender et prendre en compte la grande diversité des phénomènes observée dans la parole des patients. De telles études nécessitent le traitement de grands corpus de données, comportant suffisamment de patients, de types de dysarthrie (voire de maladies) et d'échantillons de parole pour envisager une analyse fine, rigoureuse et robuste. Dans ce contexte, une analyse manuelle seule n'est pas envisageable puisque la tâche de segmentation en phonèmes d'un signal de parole, étape préalable requise pour toute analyse phonétique, est à elle seule fastidieuse et surtout excessivement consommatrice en termes de ressources humaines. L'utilisation de systèmes issus du traitement automatique de la parole est, par conséquent, une solution à considérer dans ce contexte très particulier.

Il existe différents types de systèmes d'alignement automatique de la parole (le lecteur pourra se référer à Nefti [1] pour une revue complète). Nous nous intéresserons ici aux systèmes d'alignement contraint par le texte, qui, à partir d'une transcription orthographique du texte prononcé dans l'échantillon de parole associée à un lexique phonétisé, permettent de déterminer automatiquement les frontières des phonèmes présents. Nefti souligne que les performances de systèmes de ce type varient, dans la littérature, entre 80 et 90% de concordance avec une segmentation manuelle de référence. Par ailleurs, différents travaux ont été menés afin d'évaluer la pertinence de l'utilisation de tels systèmes dans le cadre d'analyse en phonétique et phonologie comme par exemple l'étude des voyelles [2], ou encore du schwa [3]. Cette pertinence de l'approche automatique est, dans la plupart des études, démontrée bien qu'elle soit clairement accompagnée de précautions à prendre dans l'analyse des résultats.

L'objectif de ce papier est d'étudier cette même pertinence dès lors que le système automatique est appliqué sur des échantillons de parole dysarthrique, présentant des degrés de sévérité variables. Le résultat de cette étude devrait conduire à l'élaboration de recommandations d'utilisation de ces systèmes dans ce contexte très particulier. Pour ce faire, une description du système d'alignement automatique ainsi que des procédures de corrections manuelles est présentée. Les différentes procédures et comparaisons mises en place pour mesurer la concordance entre alignement manuel et automatique sont ensuite détaillées. Enfin, l'évaluation des performances de l'alignement automatique (comparées à un alignement manuel) est faite en regard des points suivants : les types d'erreurs, les profils de locuteurs et les types de phonèmes.

2. METHODES

2.1. Procédure d'alignement automatique

L'alignement automatique utilisé dans ce travail est dit contraint par le texte dans le sens où il consiste à faire correspondre une chaîne phonétique à un signal de parole en identifiant les phonèmes produits et en segmentant les parties du signal leur correspondant (émission de frontières de début et de fin pour chaque phonème).

Développé par le Laboratoire Informatique d'Avignon (LIA), ce système repose sur l'utilisation classique de modèles de Markov Cachés (HMM) associés à un algorithme de décodage de type Viterbi (le lecteur pourra se référer à Rabiner *et al.* [4] pour une revue détaillée sur les HMM et l'algorithme Viterbi, et à Brugnara *et al.* [5] pour un descriptif du processus d'alignement). Pour accomplir sa tâche de segmentation du signal en phonèmes, le système requiert différentes ressources linguistiques : (1) une transcription orthographique « fidèle » du message linguistique véhiculé par l'échantillon de parole (prise en compte des insertions, substitutions et suppressions de mots ou sons dans le texte et des disfluences), (2) un lexique phonétisé défini à partir de la transcription orthographique, pouvant comporter, pour chaque entrée lexicale, un ensemble de variantes phonologiques, (3) un ensemble de modèles HMM représentant les différentes formes acoustiques des phonèmes estimé sur un corpus d'apprentissage. Ici, 38 modèles HMM indépendants du contexte appris sur le corpus d'émissions radiophoniques ESTER [6] sont utilisés. Pour notre étude, le lexique du système a été réduit aux lexèmes contenus dans le texte étudié. Les variantes de prononciation de ces lexèmes ont également été filtrées en fonction de ce dernier.

2.2. Procédure de correction manuelle de l'alignement automatique

La segmentation manuelle qui nous servira de référence dans cette étude a été réalisée par des experts humains à partir de l'alignement automatique. Elle consiste en une correction/vérification manuelle des étiquettes phonémiques apposées par l'aligneur automatique, et de la localisation sur le signal de leurs frontières de début et de fin. Les corrections incluent donc des ajouts, suppressions, substitutions (modifications) d'étiquettes phonémiques dans les cas où elles ne correspondent pas à ce qui a effectivement été réalisé, et des décalages temporels des frontières vers la droite ou la gauche dans les cas où les frontières placées ne correspondent pas aux critères de l'expert. S'il est difficile de définir avec certitude les frontières de phonèmes dans le continuum de parole, les phonéticiens utilisent des critères assez robustes de segmentation à partir de l'examen du signal de parole et du spectrogramme : l'apparition et la disparition du 2^{e} formant pour la segmentation des voyelles, du bruit dans les hautes et moyennes fréquences pour les fricatives, d'une tenue voisée ou silencieuse pour les occlusives (la tenue des occlusives sourdes n'est pas identifiable lorsqu'elles suivent une pause), d'un bruit correspondant au relâchement des occlusives, etc. Les aspects les plus délicats de la segmentation sont les suites de consonnes et de voyelles caractérisées par une structure de formants. En effet, le passage continu d'une articulation à une autre ne correspond pas toujours à une adresse temporelle précise.

Si la qualité essentielle d'une segmentation manuelle réside dans la consistance à toujours appliquer les mêmes critères de segmentation, il n'en reste pas moins vrai que les critères choisis sont fonction des représentations phonétiques que l'expert a du signal de parole, et peuvent donc varier d'un expert à l'autre [7]. Nous avons donc comparé les segmentations de deux experts phonéticiens sur une partie du corpus (voir plus loin pour les détails).

Dans le cas de la parole pathologique, les problèmes de segmentation sont amplifiés du fait des réalisations phonétiques souvent très perturbées des patients et de la présence de continua acoustiques quasi insegmentables. Dans ces cas, les deux experts ont eu des stratégies différentes. L'expert 1 a préféré noter ces continua comme insegmentables, tandis que l'expert 2 a tenté de distinguer les différents segments lorsque cela lui semblait possible.

2.3. Corpus

Les productions de parole dysarthrique utilisées dans cette étude ont été mises à notre disposition par l'Hôpital de la Pitié-Salpétrière. Le corpus est composé d'enregistrements de 4 patients atteints de maladies rares de surcharges lipidiques, présentant différents degrés de sévérité de dysarthrie, suivant le degré d'évolution de leur maladie. Les codes des locuteurs indiquent dans l'ordre leur genre (M ou F), le degré de sévérité de la dysarthrie (0=sujet contrôle; 1=dysarthrie légère ; 2=dysarthrie sévère), et l'initiale de leur prénom. La population, équilibrée en genre, est composée de deux patients à dysarthrie sévère (M2V et F2S) et deux patients à dysarthrie légère (M1A et F1C). Deux sujets contrôles non-dysarthriques d'âges similaires (M0A et F0D) font également partie des analyses. Nous nous sommes focalisés dans cette étude sur la lecture du texte 'Tic Tac' de la batterie de C. Chevrie-Müller. Les enregistrements ont été effectués dans un environnement calme mais non contrôlé. Les patients pouvant présenter des états de fatigue très variables, les patients dysarthriques les plus sévères ne sont pas arrivés jusqu'au bout du texte.

Les productions de ces locuteurs ont été transcrites orthographiquement selon une convention stricte et les signaux de parole ont été alignés par le système automatique. Le premier expert humain a corrigé les alignements automatiques (AA) des 4 patients, le second a corrigé les alignements des contrôles et de deux patients (M1A et F2S). Les corrections du premier expert et du second sont notées respectivement AM1 et AM2.

3. COMPARAISON DES ALIGNEMENTS AUTOMATIQUE VS. MANUEL

Les étiquettes et frontières phonémiques définies par l'AA ont été comparées à celles placées manuellement par les experts au moyen d'une procédure semi-automatique. Un premier filtrage a été nécessaire pour ne conserver que les segments comparables, excluant notamment les continua jugés insegmentables. Afin de tenir compte du décalage induit par les insertions, suppressions ou substitutions de phonèmes sur les

frontières des segments voisins, et évaluer les performances de l'AA indépendamment des erreurs issues de la transcription, les segments voisins ont également été exclus.

3.1. Divergences d'étiquettes phonémiques

Les divergences entre étiquettes proviennent en majorité des continua jugés insegmentables, dont le nombre est compris entre 40 pour M2V et 89 pour M1A dans AM1, et de 28 pour F2S et 43 pour M1A dans AM2. Les insertions, suppressions et substitutions de phonèmes sont en nombre plus restreint. Quel que soit l'alignement manuel considéré, leur nombre total est en effet compris entre 13 pour F1C et 29 pour F2S. Le nombre de segments analysés après élimination des segments non-comparables et la proportion du nombre total de segments sont présentés dans la table 1 pour chaque comparaison entre alignements.

3.2. Décalage temporel

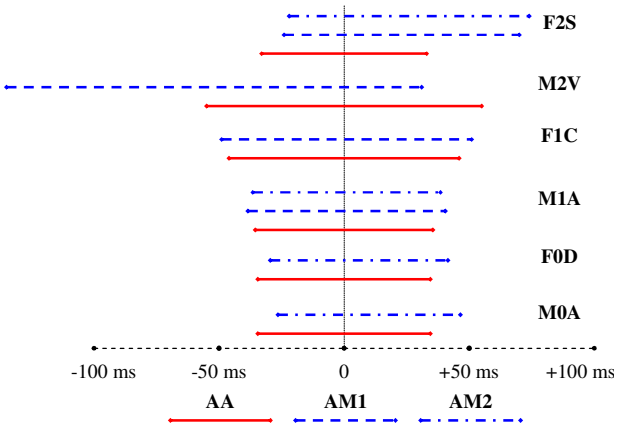

Figure 1 : Segments moyens pour chaque locuteur et chaque alignement, avec leurs décalages relatifs. Les lignes représentant les segments sont alignées sur le point central de l'AA (traits pleins rouges).

La déviation temporelle entre alignements est analysée en termes de décalage temporel entre les onsets (DecOn), points centraux (DecCtr) et offsets des unités segmentées (DecOff). Ce dernier critère cherche à vérifier si les phonèmes sont correctement alignés au signal au niveau de leur point central, même si les frontières de début et de fin sont décalées. Les valeurs négatives indiquent que le point considéré apparaît plus tôt dans le signal. La figure 1 présente les segments moyens issus des 3 alignements et leurs décalages relatifs pour chacun des 6 locuteurs. Les valeurs représentées pour les locuteurs M1A et F2S correspondent aux segments communs aux comparaisons AA vs. AM1 et AA vs. AM2. Pour chaque comparaison entre alignements, l'effet du locuteur sur les mesures de décalage a été évalué par des tests ANOVA. Les résultats de ces tests sont présentés dans la table 1. Une comparaison directe des points initiaux, centraux et finaux, dont les résultats ne sont pas détaillés ici, confirme un résultat prévisible : les alignements sont en effet tous significativement différents deux à deux.

a) Patients vs. contrôles

Il ressort de la comparaison entre AA et AM2 que la proportion de segments pour lesquels le décalage du point central est supérieur à 20 ms est peu élevé pour les locuteurs M0A (17%), F0D (14%) et M1A (21%). En revanche cette proportion est de 52% pour la locutrice F2S. Un effet significatif du locuteur sur DecCtr et DecOff est observé, les valeurs de DecOn étant en revanche comparables entre locuteurs. Les décalages sont significativement plus importants pour F2S que pour M1A et les deux sujets contrôles.

b) Décalages en fonction de la nature du segment

La répartition des décalages temporels entre AA et AM2 a ensuite été examinée par type de segment. N'ont été retenues pour l'analyse que les classes contenant au moins 10 exemplaires par locuteur. 10 classes acoustiques sont comparées : Fricatives sourdes et voisées, Occlusives sourdes et voisées, /l/, /ʁ/, Consonnes nasales, Voyelles orales, Voyelles nasales et Semi-voyelles. Les valeurs de décalage ont été examinées pour les valeurs supérieures à 20 ms, soit 2 trames pour le système automatique, sur DecCtr ou sur les différences de durée déduites de DecOn et DecOff. Cet examen révèle des décalages entre AA et AM particulièrement importants pour les occlusives sourdes, y compris pour les sujets contrôles. En effet, bien que les décalages moyens observés sur l'onset de la tenue et l'offset du burst soient généralement très faibles pour ces locuteurs (-2 à -4 ms, à l'exception de l'offset du burst pour M0A, décalé de +20 ms), indiquant un placement peu dégradé des frontières de l'occlusive considérée dans sa globalité, l'explosion est décalée de manière récurrente (-29 ms). Le locuteur M1A présente le même pattern, avec de plus l'onset de la tenue décalé de +31 ms en moyenne. La valeur moyenne de DecOn sur les semi-voyelles de M0A est de -19 ms, et celle de DecOff sur les voyelles nasales de F0D de -17 ms, pour des différences de durée respectives de +20 ms et -26 ms. Pour la locutrice F2S, des décalages importants sont observés pour l'ensemble des types de segments.

c) Décalages en fonction de la sévérité de la dysarthrie

La comparaison entre AA et AM1, qui porte sur les 4 patients, indique que la proportion de segments dont le décalage du point central est supérieur à 20 ms semble dépendre du degré de sévérité de la dysarthrie : elle est en effet de 20% pour M1A et 26% pour F1C, contre 53% pour F2S et 66% pour M2V. Un effet significatif du locuteur sur les trois mesures de décalage est mesuré, les décalages étant significativement plus importants pour M2V que pour les trois autres patients.

d) Variabilité des alignements manuels

Enfin, la comparaison entre AM1 et AM2 indique que seuls 3% des segments de M1A présentent une valeur de DecCtr supérieure à 20 ms, et 7% de ceux de F2S, tandis que pour ces 2 locuteurs le désaccord entre experts mesuré au niveau du point central est inférieur à 1 ms

dans plus de 75% des cas. Les écarts observés sont dus en grande partie à quelques segments précédés ou suivis d'un bruit de durée importante, considérés comme faisant partie du segment dans AM1 et comme distincts de ce segment dans AM2. Un effet significatif du locuteur sur les valeurs de DecDtr et DecOff est observé, mais pas sur les valeurs de DecOn.

4. DISCUSSION

L'analyse menée met en évidence une variabilité des performances du système d'alignement en fonction de la sévérité de la dysarthrie, mais également une différence entre experts humains qui, si elle n'est pas de même ampleur, n'est pas pour autant négligeable.

Une part des décalages les plus importants mesurés (jusqu'à 300 ms pour les patients les plus dysarthriques) peut s'expliquer par le fonctionnement du système d'alignement automatique. En effet, lorsque le système ne parvient pas à apparier les trames du signal avec les modèles des phonèmes supposés présents, il opère une resynchronisation de l'alignement au niveau de la portion de faible énergie suivante (modèle de silence). La majorité des trames qui n'ont pu être appariées sont alors intégrées à ce segment de faible énergie, les segments supposés présents d'après le texte à aligner étant placés au début de la portion de signal non reconnue, avec une durée de 30 ms (seuil minimal fixé par le système).

Les observations relatives au placement de l'explosion des occlusives sourdes indiquent que des mesures directement dépendantes de ces frontières comme celles de VOT ne peuvent être extraites de façon fiable à partir de l'alignement automatique. Toutefois, les décalages en cascade de segments adjacents induits par les erreurs de l'alignement automatique liés à la resynchronisation des portions de signal mal reconnues limitent la portée des autres analyses des décalages en fonction de la nature des segments.

Afin de ne pas prendre en compte dans de futures analyses acoustiques menées à partir de l'alignement automatique des segments présentant un tel décalage, la prochaine étape de notre travail sera de les filtrer automatiquement en tirant parti de mesures de confiance estimées durant le processus d'alignement automatique comme utilisé dans Clément *et al*. [8].

Remerciements : *Ce travail a été réalisé dans le cadre du projet « DesPho Apady » (ANR-08-BLAN-0125), financé par l'Agence Nationale de la Recherche (ANR).*

Nos remerciements vont à Georges Linares pour son aide sur l'utilisation du système d'alignement ainsi qu'à Nathalie Lévêque et Frédéric Sedel pour la mise à notre disposition du corpus de parole dysarthrique.

BIBLIOGRAPHIE

[1] S. Nefti. *Segmentation automatique de parole en phones. Correction d'étiquetage par l'introduction de mesures de confiance*. Thèse de doctorat, Université de Rennes 1, 2004.

[2] R. Bertrand, P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde and S. Rauzy. Le Cid - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle. *Traitement Automatique des Langues*, 49 (3):105-134, 2008.

[3] A. Bürki, C. Gendrot, G. Gravier, G. Linares and C. Fougeron. Alignement automatique et analyse phonétique : comparaison de différents systèmes pour l'analyse du schwa. *Traitement Automatique des Langues*, 49(3):165-197, 2008.

[4] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*, Prentice-Hall, 1993.

[5] F. Brugnara, D. Falavigna and M. Omologo. Automatic segmentation and labeling of speech based on Hidden Markov Models, *Speech Communication*, 12:357-370, 1993.

[6] 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 *Proceedings of LREC'06,* 2006.

[7] M. A. Pitt, K. Johnson, E. Hume, S. Kiesling and W. Raymond. The Buckeye Corpus of Conversational Speech: Labeling Conventions and a Test of Transcriber Reliability. *Speech Communication*, 45:89-95, 2005.

[8] P. Clément, C. Fredouille and N. Lévêque. Méthodes objectives appliquées à la dysarthrie, In *Actes 3èmes Journées de Phonétique Clinique*, 2009.

Table 1 : Résultats des tests ANOVA réalisés pour chaque comparaison sur les variables decOn, decCtr et decOff. Les différences entre groupes résultant des comparaisons multiples (test HSD de Tukey) sont indiquées pour p<.05. N seg : nombre et proportion de segments analysés ; N loc : nombre de locuteurs pris en compte dans la comparaison.

Comparaison	N seg	N loc	Variable	F	p	Comparaisons multiples
AA vs AM1	1319 (77%)	4	decOn decCtr decOff	11.5 7.3 4.5	<.001 <.001 .004	M2V≠(F2S, F1C, M1A) M2V≠(F2S, F1C, M1A) M2V≠(F2S, F1C, M1A)
AA vs. AM2	1917 (86%)	4	decOn decCtr decOff	2.0 35.4 9.2	.111 <.001 <.001	(F2S, M0A, F0D, M1A) F2S≠(M0A, F0D, M1A) F2S≠(M0A, F0D, M1A)
AM1 vs. AM2	716 (81%)	2	decOn decCtr decOff	0.9 10.8 20.6	.355 .001 <.001	(M1A, F2S) M1A≠F2S M1A≠F2S