

Algorithme de prédiction d'élisions de phonèmes et influence sur l'alignement automatique dans le cadre du projet Aix-MARSEC

Cyril Auran¹, Caroline Bouzon¹, Daniel Hirst¹
Christophe Lévy² & Pascal Nocéra²

¹ Laboratoire Parole et Langage, CNRS UMR 6057, Aix-en-Provence, France

² Laboratoire d'Informatique d'Avignon, France

Méls: {auran ; bouzon ; hirst}@lpl.univ-aix.fr & {christophe.levy ; pascal.nocera}@lia.univ-avignon.fr

ABSTRACT

Speech research and technology today has greater and greater need for large annotated spoken corpora. Unfortunately the number of such freely available corpora is rather limited. This paper presents the methodology and algorithms used in the generation of the Aix-MARSEC corpus, which matches these requirements. More particularly, this paper details the elision-prediction algorithm used to optimise the transcription subsequently aligned through a force Viterbi algorithm. Close analysis of the impact of the predicted elisions on the alignment is carried out. Perspectives, finally, are envisaged in order to improve the quality of the final alignment even more.

1. INTRODUCTION

Des corpus annotés et alignés représentent un intérêt considérable pour une communauté grandissante de linguistes et de spécialistes du TAL. Plus particulièrement, un alignement au niveau du phonème constitue souvent le point de départ tant pour de nombreuses études en phonétique/phonologie que pour l'entraînement de systèmes de reconnaissance automatique. Dans ce cadre, le corpus est annoté manuellement ou traité de manière automatique (reconnaissance et alignement). La première méthode représente un travail fastidieux et spécialisé impliquant de nombreuses heures de travail. La seconde méthode, notamment dans le cadre de parole spontanée, combine deux sources potentielles d'erreur (lors de la reconnaissance puis de l'alignement) et présuppose souvent un entraînement préliminaire lourd, ou de piètres résultats pour des systèmes sans adaptation au locuteur.

Cet article présente une méthodologie semi-automatique comportant une première phase de phonétisation suivie d'une phase d'optimisation de la transcription phonétique par application d'un algorithme de prédiction d'élision de phonèmes destinée à fournir une transcription plus fidèle du signal de parole. Une telle procédure constitue une alternative efficace à la reconnaissance automatique et au décodage acoustico-phonétique (DAP) et fournit une entrée plus fiable au système d'alignement.

La première partie de cet article présente la phase de phonétisation automatique ainsi que l'algorithme de prédiction des élisions. La seconde partie détaille la procédure utilisée pour l'alignement automatique de cette transcription avec le signal. La troisième partie,

finale, est centrée sur l'évaluation de l'influence de l'optimisation de la phonétisation sur la qualité de l'alignement fourni.

2. ALGORITHME DE PREDICTION D'ELISIONS

2.1. Corpus

La méthodologie présentée ici est appliquée au corpus Aix-MARSEC, version du SEC (*Spoken English Corpus*), corpus d'anglais britannique, exploitable informatiquement, composé d'enregistrements de la BBC datant des années 1980 ([1]). Les données représentent plus de 5 heures de parole naturelle transcrite orthographiquement (environ 55.000 mots) regroupées en 11 types d'enregistrement. L'évolution du SEC en MARSEC correspond d'une part à l'alignement manuel des mots et des frontières d'unités intonatives et, d'autre part, à l'annotation prosodique manuelle du corpus (G. Knowles et B. Williams) fondée sur un ensemble de marques tonétiques.

2.2. Phonétisation

Dans ce travail, les étiquettes au niveau du mot ont été vérifiées et corrigées manuellement avant que les mots ne soient automatiquement phonétisés à partir d'un dictionnaire électronique (*Advanced Learners' Dictionary*) converti en alphabet SAMPA [2]. La phonétisation, effectuée à l'aide de scripts Perl, s'appuie sur ce dictionnaire ainsi que sur une liste de 700 mots transcrits manuellement (absents du dictionnaire). Une étape ultérieure consiste à traiter les réductions de mots outils (tels que "a", "the", etc.) à l'aide d'une liste de mots réduits dès lors qu'ils ne comportent pas de marque tonétique. Les nombres et les dates reçoivent un traitement spécifique (conversion en mots puis phonétisation). Les abréviations alphabétiques (de type "BBC") sont décomposées en lettres avant d'être phonétisées. Les abréviations mixtes ("BBC2") et les codes postaux ("Y01 IET") reçoivent un traitement qui sépare les nombres des lettres, transcrit les nombres en mots et phonétise la séquence « mots-lettres ». Le résultat de cette phase globale de phonétisation constitue ce que nous appelons la « transcription brute ».

2.3. Optimisation de la phonétisation

La transcription brute est ensuite optimisée à l'aide de règles d'élision de phonèmes obtenues à partir de

l'observation d'échantillons du corpus et de [3], [4] et [5]. L'objectif est d'améliorer la transcription et la qualité globale de l'alignement automatique (augmentation de la correspondance entre signal et transcription).

L'hypothèse d'« élasticité » de Campbell ([6]) est étendue au niveau du mot, menant au calcul d'un *facteur* z correspondant à la *transformée* z (cf. (formule 1)).

$$Durée_Mot = \sum_{i=1}^{Nb_phonèmes} (moy_{t_{phoi}} + z * e_{t_{phoi}}) \quad (1)$$

Une valeur négative pour le facteur z peut alors être interprétée comme un raccourcissement relatif du mot, qui constitue une base solide pour l'application des règles d'élimination, fondées à la fois sur des critères phonotactiques et un ensemble de seuils de durées minimales pour certains phonèmes (cf. [7] : 289).

Lors de l'application des règles, 4077 phonèmes sont éliminés de la transcription brute, menant ainsi à ce que nous appelons la « transcription optimisée ».

3. PROCEDURE D'ALIGNEMENT

Cet article fait appel à deux approches d'alignement en phonème :

- Un système entièrement automatique fondé sur un DAP ;
- Un système semi-automatique nécessitant une transcription phonétique.

La méthode utilisant un DAP permet de générer automatiquement la transcription phonétique ; elle sera comparée à la transcription ayant subi les élisions et à la transcription manuelle (cf. 4.1.). Le système semi-automatique, pour sa part, est utilisé pour l'alignement des transcriptions brutes et des transcriptions optimisées. Les deux méthodes (DAP et semi-automatique) sont fondées sur une approche commune : les phonèmes sont modélisés par des Modèles de Markov Cachés (MMC [8]) et sont ensuite décodés avec l'algorithme de Viterbi ([9]).

3.1. Modèles de Markov Cachés

Comme classiquement en reconnaissance automatique de parole continue, les phonèmes non-contextuels sont modélisés par des MMCs, gauche-droite, à trois états émetteurs. Les modèles sont appris sur les données du corpus TIMIT (630 personnes, parlant 8 des principaux dialectes de l'anglais américain, qui prononcent chacune dix phrases) contenant 48 phonèmes.

Les corpus que l'on cherche, généralement, à aligner sont inconnus, ce qui implique que les modèles de phonèmes sont appris sur des corpus différents de celui que l'on cherche à aligner. Il est donc nécessaire, comme première étape, d'apprendre des modèles avec des corpus alignés manuellement.

Chaque état des MMCs est composé d'un mélange de 8 gaussiennes (représentées par leur moyenne et la

diagonale de la matrice de covariance). Ces paramètres sont estimés grâce à l'algorithme E-M (Expectation-Maximisation) en optimisant le critère du maximum de vraisemblance.

Le signal de parole est paramétrisé en 12 coefficients MFCCs (Mel Frequency Cepstral Coefficients) obtenus suite à une analyse en banc de filtre. Ces 12 coefficients sont complétés par l'énergie du signal. Enfin, les dérivés de premier et second ordres sont calculés et concaténés avec le vecteur initial afin d'obtenir un vecteur de 39 coefficients pour chaque trame de signal.

3.2. DAP vs. Alignement forcé

Les deux méthodes (DAP et semi-automatique) sont fondées sur le même algorithme de décodage : Viterbi. Cet algorithme calcule la séquence optimale (la plus vraisemblable) étant donnée une série d'observations.

Lors d'un DAP, la suite de phonèmes n'est pas contrainte : chaque phonème peut apparaître avec la même probabilité. La suite phonétique décodée est obtenue en maximisant le critère de vraisemblance. Cet algorithme donne, *in fine*, la séquence phonétique la plus probable, ainsi que les limites inter-phonèmes.

Avec l'alignement forcé, la séquence de phonèmes est pré-définie, c'est-à-dire que les transitions entre phonèmes sont fixées. Dans ce cas, l'algorithme de Viterbi est uniquement utilisé pour affecter les trames aux états, donc pour déterminer les limites inter-phonèmes (comme lors du DAP).

4. EVALUATION

L'évaluation de l'algorithme présenté ici se divise en deux parties. La première partie consiste à évaluer la fiabilité d'un DAP et celle de notre algorithme de phonétisation par dictionnaire (phonétisation optimisée) par rapport à une transcription et un alignement manuels. La seconde partie porte sur l'évaluation de l'influence de l'algorithme de prédiction d'élisions sur la qualité de l'alignement lui-même.

Ces deux étapes, en l'état actuel, utilisent comme référence la transcription et l'alignement manuels de 4 fichiers extraits du corpus (ce qui représente approximativement 4 minutes de parole). L'alignement et l'évaluation de près d'une heure de ce même corpus sont en cours, permettant ainsi d'augmenter la significativité des résultats présentés ici.

4.1. Phase de transcription : DAP vs. phonétisation optimisée

La première phase de cette évaluation quantifie la fiabilité de deux méthodes de phonétisation (DAP vs. phonétisation optimisée). Dans cette perspective, notre centre d'intérêt porte sur l'identité des étiquettes (nature de chaque étiquette de phonème dans la transcription considérée). La fiabilité de l'alignement (correspondance avec les frontières manuelles de chaque phonème), étant

directement liée à l'utilisation du même décodage Viterbi pour les deux transcriptions, ne sera pas détaillée ici.

L'évaluation automatique des transcriptions a été effectuée à l'aide d'un script Perl et peut être résumée ([10]) à l'aide des valeurs de précision (table 1).

Table 1 : Précision des deux transcriptions

Transcription	Précision
DAP	28.97 %
Transcription optimisée	94.79 %

Les deux systèmes évalués ici constituent des tâches de complexités fort différentes (le DAP constituant une tâche plus complexe), dont la comparaison peut sembler délicate. On notera cependant que la différence considérable observée ici entre la précision du DAP (habituellement située dans une fenêtre 60-70%) et celle de la phonétisation optimisée peut recevoir deux explications principales :

- Des contraintes pratiques de disponibilité nous ont conduits à devoir utiliser des modèles phonétiques entraînés sur le corpus TIMIT d'anglais américain ; de manière évidente, de tels modèles, présentent des différences significatives avec le standard britannique du corpus Aix-MARSEC.

- La durée moyenne des fichiers son (environ 1 minute) a impliqué un fractionnement en unités plus petites (unités intonatives définies dans le SEC), processus ayant pu influencer le DAP de manière significative au niveau des frontières ainsi créées.

Cette phase d'évaluation montre que, bien qu'une méthode de DAP puisse être utilisée, une méthode de phonétisation fondée sur la méthodologie présentée ici constitue une alternative intéressante.

4.2. Elisions et fiabilité de l'alignement

De manière évidente, une phonétisation plus fiable améliore la qualité globale de l'alignement, la correspondance entre la transcription et le signal de parole étant plus élevée. Il est alors crucial dans ce cadre de quantifier l'influence de notre algorithme de prédiction d'élosion de phonèmes sur l'alignement, notamment afin de vérifier qu'aucun biais n'est introduit.

Pour tester cette hypothèse, l'alignement automatique des quatre fichiers utilisés dans cette étude, fondé sur la transcription brute ou sur la transcription optimisée, a été systématiquement comparé à l'alignement manuel de la transcription considérée. La sortie de l'alignement automatique de la transcription brute a ainsi été comparé avec l'alignement manuel de cette même transcription ; de manière similaire, l'alignement automatique de la transcription optimisée a été comparé avec l'alignement manuel de cette même transcription. Plus précisément, les phonèmes donnés dans la transcription mais absents du signal ont été réduits à 10ms afin de faciliter la comparaison avec l'alignement automatique qui, fondé sur une fenêtre d'analyse de 10ms, fonctionne de manière similaire. Ceci explique les résultats parfois

apparemment meilleurs observés avec la transcription brute (cf. ligne « 20 ms » (table 2)). La (table 2) détaille la qualité de l'alignement en fonction de la transcription utilisée (brute vs. optimisée) et du seuil d'acceptation (SA) [11].

Table 2 : Qualité de l'alignement exprimée en pourcentage d'étiquettes alignées à moins d'un SA donné par rapport à l'alignement manuel.

SA	Transcription brute	Transcription optimisée
64 ms	93.03 %	93.25 %
32 ms	82.02 %	82.02 %
20 ms	68.92 %	68.37 %
16 ms	59.99 %	59.97 %
15 ms	57.43 %	57.40 %
10 ms	42.21 %	42.43 %
5 ms	23.63 %	23.72 %

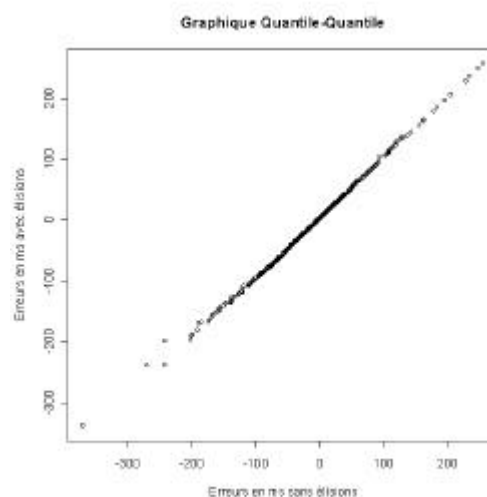


Figure 3 : Quantile-Quantile des distributions d'erreurs d'alignement pour les transcriptions brute et optimisée.

La (figure 3) confirme graphiquement l'hypothèse de la similarité entre les distributions des erreurs d'alignement fondé sur les transcriptions brute et optimisée. En effet, le graphique prend la forme d'une droite à 45°, typique d'une correspondance quasi parfaite entre les échantillons. Des tests formels, cependant, doivent être effectués afin de corroborer cette assertion.

Plusieurs procédures statistiques peuvent être employées pour tester de manière formelle l'absence de biais dans l'alignement dû à l'optimisation de la transcription. Les distributions des erreurs, cependant, divergent de manière significative de distributions normales (table 3), rendant ainsi inadaptées les t-test et F-test classiques.

Table 3 : Aplatissement et dissymétrie des distributions des erreurs en ms.

	Transcription brute	Transcription optimisée
Aplatissement	13.07	14.64
Dissymétrie	-0.29	-0.57

Des tests de comparaison formels ont été effectués à l'aide de l'environnement R ([12]):

- Le test de sommes ordonnées avec correction de continuité de Wilcoxon (qui ne présuppose pas la normalité des échantillons) fournit une *p-value* de 0.7757, qui confirme l'hypothèse de l'égalité des moyennes pour les deux échantillons considérés.
- Le test de Kolmogorov-Smirnov, avec une *p-value* de 1, confirme lui aussi sans équivoque que malgré des tailles d'échantillons nécessairement différentes, les deux alignements ne différencient pas de manière significative.

Ces résultats confirment l'absence de biais dû à l'application de l'algorithme de prédiction d'élosion de phonèmes. La qualité globale de l'alignement dépend directement de la fiabilité de la transcription phonétique générée dans le cadre de notre méthodologie. Avec une précision d'environ 95% (§4.1.), la méthodologie Aix-MARSEC semble ainsi constituer une base solide en vue de la phonétisation et de l'alignement de grandes bases de données orales.

Il est à noter, finalement, que l'évaluation de la qualité de l'alignement pour les mots comportant une élosion est en cours, et permettra de mettre en évidence de manière plus claire l'impact de notre démarche sur l'alignement des segments de ce domaine plus restreint.

5. CONCLUSION ET PERSPECTIVES

Cet article argumente en faveur d'une méthodologie de phonétisation impliquant des règles de prédiction d'élosion de phonèmes. En effet, une telle démarche fournit des résultats significativement meilleurs que ses contre-parties totalement automatiques, que ce soit pour la nature de la transcription ou pour l'alignement de cette dernière sur le signal.

Plusieurs perspectives d'amélioration sont envisagées :

- Après évaluation détaillée ([13]), la première version de l'algorithme prédit des élosions avec une précision de 74.44%, une F-mesure de 60.18%, et un silence de 49.49% ; cette dernière mesure reflète le fait que près de la moitié des élosions effectivement réalisées par les locuteurs n'est pas, en l'état, prédite. De nouvelles règles d'élosion (pour certaines proposées automatiquement lors de l'évaluation) seront introduites dans l'algorithme afin d'améliorer la transcription optimisée.
- Étant donné que la procédure d'alignement fixe une durée minimale de 10ms pour les phonèmes non détectés, la qualité de la transcription pourrait être améliorée par une méthode itérative contrainte d'élosion de ces phonèmes (contraintes phonotactiques notamment).
- L'utilisation d'une transcription marquant les sites potentiels d'élosion de phonème (prédits par notre algorithme) est également envisagée, permettant de finaliser la transcription phonétique lors de la phase d'alignement.
- Une méthode alternative d'alignement, fondée sur l'utilisation itérative d'un algorithme de programmation dynamique et s'appuyant sur la comparaison d'un signal

de synthèse avec le signal d'origine, reste à être implémentée et évaluée ([11]).

- Finalement, la correction manuelle de l'alignement de près d'une heure (actuellement effectué par les étudiants de DEA du *English Prosody Group of Aix*) permettra l'entraînement de modèles phonétiques spécifiques permettant d'améliorer l'alignement de la totalité du corpus et d'évaluer l'apport de notre démarche sur une quantité plus significative de données.

Aligné au niveau du phonème (eux-mêmes regroupés automatiquement en syllabes, constituants sub-syllabiques, pieds accentuels, mots et unités intonatives), le corpus Aix-MARSEC constitue une base de données d'anglais britannique oral, disponible librement en contactant l'un des auteurs (www.lpl.univ-aix.fr/~EPGA/fr_marsec.html).

6. BIBLIOGRAPHIE

- [1] G. Knowles, A. Wichmann and P. Alderson. *Working with Speech: perspectives on research into the Lancaster/IBM Spoken English Corpus*. England: Longman, 1996.
- [2] SAMPA, disponible sur le site: <http://www.phon.ucl.ac.uk/home/sampa/home.htm>
- [3] J.C. Wells. *Pronunciation Dictionary*. England: Longman, 1990.
- [4] D. Jones. *English Pronouncing Dictionary*. England: Longman, 1991.
- [5] A. Cruttenden. *Gimson's Pronunciation of English. Fifth edition*. England: Arnold, 1997.
- [6] N. Campbell. *Multi-level Timing in Speech*. PhD Thesis, University of Sussex, 1992.
- [7] D.H. Klatt. Synthesis by Rule of Segmental Durations in English Sentences. In B. Lindblom and S.E.G. Ohman (eds): *Frontiers of Speech Communication Research*, pages 287-299, 1979.
- [8] L.R. Rabiner. A tutorial on hidden Markov Models and selected applications in speech recognition, *IEEE transactions on Speech Audio Processing*, vol. 2, 1984.
- [9] A. Viterbi. Error bounds for convolutional Codes and an asymptotically optimum decoding algorithm, *IEEE Transactions on Information Theory*, vol. 2, 1967.
- [10] C.J. Van Rijsbergen. *Information Retrieval, 2nd edition*. University of Glasgow, 1979.
- [11] P. Di Cristo et D.J. Hirst. Un procédé d'alignement automatique de transcriptions phonétiques sans apprentissage préalable. *4° Congrès Français d'Acoustique*, 1, Marseille, 14-18 avril, France : SFA, Teknea, 1997.
- [12] The R Project for Statistical Computing, disponible sur le site: <http://www.r-project.org>
- [13] C. Auran and C. Bouzon. Phonotactique prédictive et alignement automatique : apports et perspectives pour le traitement de grands corpus oraux, *TIPA*, volume 22, 2004.