Gestion d'erreurs pour la fiabilisation des retours automatiques en apprentissage de la prosodie d'une langue seconde

Anne Bonneau — Dominique Fohr — Irina Illina — Denis Jouvet — Odile Mella — Larbi Mesbahi — Luiza Orosanu

Speech Group[1]Inria, 615 rue du Jardin Botanique, Villers-lès-Nancy, F-54600, France

Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54506, France

CNRS, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54506, France

{prénom.nom}@loria.fr

RÉSUMÉ. *Le succès des futurs systèmes d'apprentissage de l'oral d'une langue seconde assisté par ordinateur repose sur la fourniture à l'apprenant de diagnostics personnalisés et de corrections pertinentes de ses prononciations. Après une présentation de la problématique des retours prosodiques automatiques fiables en apprentissage des langues, nous présentons nos travaux sur la prise en compte de certaines erreurs provenant de l'apprenant et du système. Ainsi, la première partie concerne le rejet des entrées incorrectes (par exemple dues à des erreurs de l'apprenant) tout en étant tolérant aux déviations de la parole non native. La seconde partie porte sur la segmentation phonétique automatique de la parole non native. Une analyse détaillée a montré l'apport de la prise en compte des variantes non natives, et a permis de déterminer les classes de phonèmes dont les frontières temporelles sont les plus précises et qu'il faudra privilégier dans la conception des exercices pour l'apprentissage.*

ABSTRACT. *The success of future syste ms for computer assisted foreign language learning relies on providing the learner personalized diagnosis and relevant corrections of its pronunciations. After a presentation of the problem of reliable automatic prosodic feedbacks in language learning, we present our work related to the processing of some errors stemming from the learner and from the system itself. The first part deals with the relevant rejection of incorrect entries (for example due to learner's errors) while being tolerant to non-native speech deviations. The second part focuses on the automatic phonetic segmentation of non-native speech. A detailed analysis has showed the benefit of taking into account non-native variants, and lead to determining the classes of phonemes whose temporal boundaries are the most accurate and which should be favored in the design of exercises for language learning.*

MOTS-CLÉS : *apprentissage des langues, parole non native, alignement texte-parole, segmentation phonétique, modélisation stochastique de la parole, rejet entrées incorrectes.*

KEYWORDS: *language learning, non-native speech, text-to-speech alignment, phonetic segmentation, stochastic modeling of speech, rejection of incorrect entries.*

1. Introduction

Tout le monde s'accorde aujourd'hui sur la nécessité d'apprendre au moins une langue étrangère. Mandatés par le Conseil européen de Lisbonne (2000), les ministres de l'Éducation font de l'enseignement des langues une priorité et insistent sur l'importance de se former tout au long de sa vie (processus de Lisbonne[1]). La nécessité de cet apprentissage et ce qu'il est convenu d'appeler l'essor des nouvelles technologies ont conduit à de nombreuses recherches et expérimentations (Speech Communication, 2009 : Eskenazi, 2009) et ont permis l'apparition sur le marché de nombreux logiciels d'apprentissage assisté par ordinateur. Pourtant, après des débuts encourageants, les logiciels commerciaux ont fait face à un certain manque d'intérêt. La raison de ce recul vient probablement à la fois du diagnostic effectué et du retour d'information vers l'apprenant, tous deux très limités, et de l'absence de fiabilité des logiciels proposés. Pour que de véritables retours fondés sur les productions des apprenants puissent être mis en œuvre dans un système d'apprentissage assisté par ordinateur, il est nécessaire de disposer de systèmes de reconnaissance de la parole et de détection d'événements acoustiques fiables.

Bien que les systèmes de reconnaissance automatique de la parole aient maintenant atteint un niveau de performance suffisant pour être utilisés dans des applications spécifiques telles que la dictée vocale, par exemple Dragon Naturally Speaking (NUANCE[2]), la commande et le contrôle de systèmes GPS (MAGELLAN[3], TOM-TOM[4]), ou encore pour l'accès à des centres d'appels (LOQUENDO[5], NUANCE[6]), la reconnaissance fiable de la parole spontanée pour n'importe quel locuteur, et la reconnaissance de locuteurs non natifs ayant un fort accent étranger constituent encore des défis. En dehors des technologies de reconnaissance de la parole, d'autres traitements automatiques de la parole sont nécessaires pour l'apprentissage de la prosodie d'une langue, comme par exemple la détection de la courbe mélodique ; ainsi, la détection de la fréquence fondamentale suscite toujours de nouvelles recherches (Zahorian et Yu, 2008) et sa fiabilité constitue encore un défi pour des voix d'enfants.

En outre, quand les technologies de traitement automatique de la parole sont appliquées à l'apprentissage des langues assisté par ordinateur (ALAO), il devient nécessaire de gérer différents types d'erreurs : d'une part, celles commises par le locuteur, qui doivent être détectées et donner lieu à un retour d'information pertinent et pédagogique, et, d'autre part, celles commises par le système lui-même, qui doivent être maîtrisées. Les réponses les plus appropriées à ce type de problème

1. www.ciep.fr/sitographie/ries46.php
2. www.nuance.com/naturallyspeaking/
3. www.magellangps.com/products/voice.asp
4. www.tomtom.com/news/category.php?ID=4&NID=368&Lid=1
5. www.loquendo.com/en/technology/speechsuite.htm
6. www.nuance.com/for-business/by-solution/contact-center-customer-care/

passent probablement par une conception intelligente des activités d'apprentissage et la mise au point d'outils dédiés à la gestion des erreurs. En effet, une bonne gestion des erreurs est cruciale dans les domaines de l'apprentissage puisque l'impact de corrections erronées, fussent-elles peu fréquentes, peut être très néfaste en apprentissage.

Des retours automatiques fiables sur la prosodie, incluant notamment un retour explicite provenant de diagnostics sur la durée et sur la ligne mélodique, ainsi qu'un retour perceptif implicite provenant de modifications de la parole (Henry *et al.*, 2007 ; Bonneau et Colotte, 2011), sont un des objectifs du projet ALLEGRO[7] auquel nous participons. Il est important de préciser que le diagnostic et les modifications de la parole sont faits à partir d'un énoncé court : le locuteur prononce une phrase ou un mot et reçoit un retour immédiat. Avec ce type de fonctionnement interactif, qui apparaît comme très bénéfique pour l'apprenant (Marty, 1983), une erreur de segmentation peut conduire à une correction erronée, ce qui doit être évité dans la mesure du possible. Un bilan prosodique portant sur un grand énoncé et ne corrigeant pas des points précis serait moins risqué mais également moins bénéfique à l'apprenant. Ce papier décrit les méthodes que nous avons mises au point dans le cadre du projet ALLEGRO et qui sont destinées à limiter et maîtriser les risques survenant lors de l'acquisition et de la segmentation de la chaîne sonore. Les retours automatiques sur la prosodie ne sont pas l'objet de cet article, mais ils sont évoqués ici dans la mesure où ils expliquent l'utilité des méthodes développées et présentées ici.

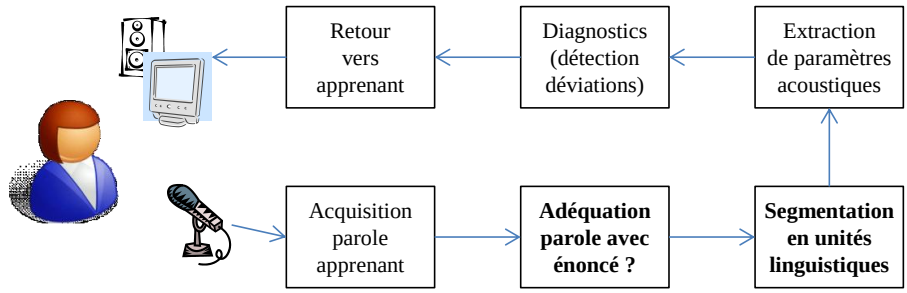


Figure 1. *Principaux modules d'un système automatique pour l'apprentissage de la prosodie d'une langue seconde assisté par ordinateur*

Un système automatique pour l'apprentissage de la prosodie d'une langue seconde assisté par ordinateur doit effectuer un ensemble de traitements qui partent de l'enregistrement d'une prononciation d'un apprenant pour aller jusqu'à la fourniture des retours (figure 1), et qui comportent tous des risques d'erreurs. Ces

7. www.allegro-project.eu/

traitements comprennent notamment l'acquisition du signal de parole de l'apprenant (c'est-à-dire l'enregistrement, la numérisation d'une prononciation non native, et la détection du segment de parole utile), la vérification de l'adéquation du signal vocal par rapport à l'énoncé attendu (aussi dénommée rejet des entrées incorrectes), la segmentation des réalisations en unités linguistiques (sons, syllabes, mots...), l'extraction de paramètres acoustiques (par exemple la ligne mélodique pour la prosodie), l'établissement de diagnostics à partir d'analyses acoustico-phonétiques ou d'une reconnaissance automatique qui décèlent les déviations de l'apprenant, et enfin, l'élaboration et la transmission des retours sous forme visuelle ou auditive.

Les méthodes de fiabilisation développées dans le cadre du projet ALLEGRO portent sur le rejet des entrées incorrectes (prononciations ne correspondant pas au texte attendu) ainsi que sur la segmentation en unités linguistiques. Les erreurs liées au bruit, au sens acoustique du terme, et à la détection de la parole seront prises en compte ultérieurement.

Dans le domaine de l'apprentissage de la prosodie, la visualisation des courbes mélodiques de l'apprenant, en général accompagnée d'une ligne mélodique de référence – issue d'un modèle ou provenant d'un locuteur natif – a été proposée dès le début des années 60, que ce soit pour l'apprentissage d'une langue seconde ou pour aider les déficients auditifs. Vardanian (1964) a été l'une des premières à utiliser et tester l'apport de cette visualisation pour des apprenants d'une langue seconde. Après une période de scepticisme, les résultats de James (1977) fondés sur les analyses du détecteur de la fréquence fondamentale de Philippe Martin (Germain-Rutherford et Martin, 2000), ou ceux de de Bot (1983), démontrent l'efficacité de ces retours. Mais la simple visualisation des courbes mélodiques, qu'elles soient accompagnées ou non des retours auditifs classiques, si elle apparaît intéressante, ne donne pas aux apprenants de retour explicite sur leur production, et, comme Chun (1998) le remarque, ceux-ci doivent « extrapoler » eux-mêmes leurs déviations. Une aide plus efficace consisterait donc à fournir aux apprenants des indications sur ces déviations et sur la manière de s'améliorer. Ce diagnostic automatique doit être fondé sur une véritable analyse acoustico-phonétique de la production de l'apprenant qui ne peut être effectuée qu'à partir d'une segmentation en unités (sons ou syllabes) précise. Par exemple, une segmentation précise est nécessaire pour juger des durées relatives des syllabes accentuées ou de leur noyau vocalique par rapport à celles des syllabes adjacentes, un élément important pour estimer la réalisation de l'accent lexical. La segmentation est également nécessaire pour juger du contour mélodique associé à l'accent de type « focus », ou des contours montants apparaissant en fin de phrases intonatives pour certains types de questions. Or les systèmes de reconnaissance automatique de la parole fondés sur les modèles de Markov cachés, qui sont les modèles également utilisés pour la segmentation automatique, ne nécessitent pas, pour une grande partie de leurs applications (comme la dictée automatique), une segmentation temporelle précise, et la précision des frontières n'est pas un critère explicite ni un critère implicite lors de l'apprentissage des modèles acoustiques. Améliorer la précision temporelle de la

segmentation est donc une des tâches prioritaires pour l'amélioration des systèmes d'apprentissage automatique de l'oral d'une langue assisté par ordinateur. En apprentissage des langues, pour de nombreux exercices de prononciation, le texte est connu à l'avance, et la segmentation automatique se fait alors par alignement texte-parole. L'alignement texte-parole est délicat pour de la parole non native, en particulier à cause des nombreuses variantes de prononciation non natives qu'il est indispensable de considérer. De plus, la segmentation entre deux sons est parfois relativement aisée, et d'autres fois très délicate, que ce soit par un expert humain ou par un système automatique. Une bonne connaissance des frontières qui peuvent être mises de manière fiable doit permettre d'adapter les exercices d'apprentissage afin de pouvoir apporter des retours adéquats et fiables.

Cet article est organisé de la manière suivante. La section 2 présente les caractéristiques des corpus utilisés lors de cette étude. Les sections suivantes décrivent la prise en compte de trois types d'erreurs. La section 3 détaille le rejet des entrées incorrectes, qui correspondent aux phrases pour lesquelles le locuteur n'a pas prononcé le texte demandé. La section 4 traite de la robustesse de l'alignement phonétique automatique face aux déviations de prononciation non natives (remplacements, omissions ou insertions des sons), et analyse, entre autres, l'apport de la prise en compte de variantes de prononciation non natives lors de l'alignement texte-parole. La section 5 est consacrée aux erreurs de la segmentation phonétique automatique ; elle analyse et commente en particulier la précision des frontières des phonèmes en fonction des classes de sons.

2. Corpus

L'évaluation des performances de rejet des entrées incorrectes et de la qualité de l'alignement phonétique automatique a été effectuée sur un corpus de parole non native, en l'occurrence un corpus de phrases anglaises prononcées par des locuteurs français. De plus, comme les approches développées reposent sur une modélisation statistique de la parole (modèles de Markov cachés), deux corpus de parole native ont été utilisés pour l'apprentissage des modèles acoustiques : un corpus de parole en anglais américain prononcé par des américains, et un corpus de parole en français prononcé par des français.

TIMIT (Garofolo *et al*., 1993) est un corpus de parole lue en anglais américain. Il contient les enregistrements de 630 locuteurs américains, chacun lisant dix phrases phonétiquement riches. Il a servi pour apprendre les modèles acoustiques des phonèmes de l'anglais et du silence.

ESTER2 est un corpus d'émissions radiophoniques (environ 180 heures), issu de la campagne d'évaluation ESTER2 (Galliano *et al*., 2009). Il a servi pour apprendre les modèles acoustiques des phonèmes du français, ainsi qu'un modèle acoustique de bruit. Ces modèles ont été utilisés pour la prise en compte de variantes de

prononciation non natives qui faisaient intervenir des phonèmes du français qui n'existent pas en anglais, comme par exemple le schwa, /y/ et /ɑ̃/.

Le corpus de parole non native provient du projet INTONALE (Dargnat *et al.*, 2010). La partie utilisée correspond à un total d'environ huit cents phrases anglaises prononcées par trente-quatre locuteurs français (vingt-neuf femmes et cinq hommes), soit environ vingt-trois phrases par locuteur. Les enregistrements ont été effectués dans une pièce calme. Lors de l'enregistrement, le texte des phrases à prononcer était affiché sur l'écran, et, après chaque prononciation d'une phrase, le locuteur pouvait choisir de la répéter (en cas de problème) ou sinon de passer à la phrase suivante.

La longueur moyenne des phrases est de huit mots (un mot pour la plus courte, quinze mots pour la plus longue). La durée de la parole varie de 1 seconde à 7 secondes, avec une durée moyenne de 2,7 secondes. Les histogrammes de la figure 2 précisent la répartition des longueurs des transcriptions des énoncés (en mots) et des durées de parole des énoncés (en secondes). Ce corpus non natif correspond à une totalité de 6 643 occurrences de mots (pour 167 entrées lexicales différentes), et environ 35 minutes de signal de parole. Les transcriptions orthographiques originales (celles affichées sur l'écran) ont été ensuite manuellement corrigées lorsque c'était nécessaire de manière à refléter le contenu linguistique de la phrase telle que prononcée par le locuteur.

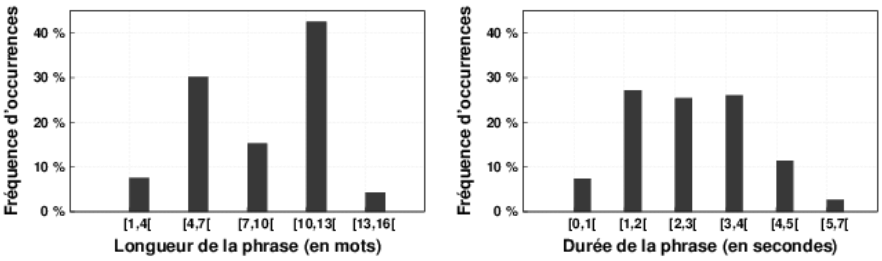


Figure 2. *Fréquence d'occurrences des longueurs des phrases (en mots, à gauche) et des durées de parole des énoncés (en secondes, à droite)*

Le corpus de parole non native a été annoté phonétiquement en deux étapes. D'abord, un alignement phonétique a été réalisé automatiquement à partir d'un alignement texte-parole avec un lexique de prononciations correspondant aux variantes de prononciation natives. Ensuite la segmentation phonétique a été corrigée manuellement ; la consigne principale était de bien vérifier et corriger les frontières des segments phonétiques, mais il n'était pas demandé de corriger précisément le timbre des voyelles. Le corpus a été divisé en deux parties d'environ 400 phrases chacune, une partie (apprentissage) est utilisée pour la détermination

des variantes de prononciation non natives et l'adaptation des paramètres des modèles, et l'autre (test) est réservée pour les évaluations de performances.

Le tableau 1 donne quelques exemples de phrases prononcées. Les variantes de prononciation non natives seront présentées et discutées dans la section 4, quelques exemples issus du corpus seront fournis à cette occasion (tableau 2).

The pupils.
Does he look at Maria?
I've got only her cell phone number.
I thought you'd like it. I found it very interesting.
Great, the weather was fine. I visited a lot of natural sites.

Tableau 1. *Exemple de phrases du corpus de parole non native*

3. Détection des entrées incorrectes

La détection des entrées incorrectes vise à identifier les entrées – couple signal de parole acquis et texte de la phrase attendue – qui sont incorrectes, c'est-à-dire celles pour lesquelles le signal de parole ne correspond pas au texte attendu, soit à cause d'une inattention de l'apprenant, soit en raison de la présence de paroles parasites ou de la présence de bruits. Dans cette section, nous nous intéressons essentiellement aux deux premiers cas, *i.e.* une incohérence au niveau des mots (mots manquants, en trop ou remplacés par d'autres). Pour une entrée incorrecte, il n'est pas possible d'établir un retour détaillé pertinent vers l'apprenant ; la réponse la plus appropriée est une demande de répétition, ou un retour rappelant simplement la phrase attendue, voire un message *ad hoc*. Ce type de détection s'apparente au rejet des entrées incorrectes ou des mots hors vocabulaire en reconnaissance de la parole (Boite, 2000).

Les méthodes de détection de mots hors vocabulaire proposées dans la littérature peuvent être classées en deux catégories. La première approche consiste à modéliser les mots hors vocabulaire soit avec des modèles de fragments, *i.e.* avec des unités sous-lexicales (Bisani et Ney, 2005), soit avec des modèles « poubelles ». La seconde approche consiste à détecter les mots hors vocabulaire en utilisant par exemple des probabilités *a posteriori* ou des mesures de confiance (White *et al.*, 2008 ; Lecouteux et Linarès, 2009). Le problème de détection et de rejet des entrées incorrectes intervient également dans le domaine de la détection de mots ou d'expressions clés (*Spoken Term Detection*) pour l'indexation et la recherche d'information dans de larges archives vocales. Pour compenser les erreurs de reconnaissance phonétique lors de la constitution de l'index phonétique des données, un modèle probabiliste des prononciations est proposé dans (Pinto *et al.*, 2008) qui est obtenu à partir des prononciations des mots et d'une matrice de confusion entre

phonèmes. Dans (Davel *et al*., 2012), la comparaison de séquences de phonèmes et une approximation de la probabilité phonétique *a posteriori*, ont été exploitées dans un contexte de validation de corpus de parole collecté à partir de smartphones.

Dans notre travail, l'approche choisie pour la détection des entrées incorrectes repose sur la comparaison de différents alignements phonétiques résultant de décodages plus ou moins contraints : un alignement contraint, où l'on force le décodage à respecter la séquence des mots de la phrase attendue (*i.e.* alignement standard texte-parole), et deux alignements non contraints issus, pour l'un, d'un décodage fondé sur une boucle de mots, et pour l'autre, d'un décodage fondé sur une boucle de phonèmes (Orosanu *et al*., 2012). La boucle de mots repose sur le vocabulaire utilisé dans le corpus, soit environ 170 mots. Pour les deux décodages non contraints, les unités (mots ou phonèmes selon le cas) sont équiprobables. Chacun de ces trois décodages fournit une segmentation phonétique de l'énoncé. Les critères calculés pour le rejet résultent de la comparaison de l'alignement contraint avec chacun des alignements non contraints (*i.e.* alignements résultant des décodages non contraints). Ces informations sont ensuite combinées et une décision d'acceptation ou de rejet est prise.

3.1 *Critères pour la détection automatique des entrées incorrectes*

La séquence de phonèmes fournie par un alignement non contraint correspond, aux erreurs de reconnaissance près, à la suite des sons prononcés par l'apprenant. En revanche, la séquence de phonèmes fournie par un alignement contraint correspond très bien aux sons prononcés par l'apprenant lorsque l'entrée est correcte (et en adéquation avec une des variantes de prononciation prévues), mais ne correspond pas lorsque l'entrée est incorrecte (ou non conforme aux variantes prévues). Les critères calculés visent à faire ressortir cette différence de comportement de l'alignement contraint au niveau des phonèmes, des trames, des segments de non-parole, des rapports de vraisemblance et des durées des phonèmes. Les alignements contraints tiennent compte de variantes de prononciation non natives.

3.1.1. *Critère associé aux phonèmes*

Comme le montre la figure 3, ce critère mesure l'adéquation des suites de phonèmes entre les alignements contraints et non contraints. Plus précisément, il s'agit du pourcentage de segments phonétiques qui ont la même étiquette dans les deux segmentations, et dont au moins une limite temporelle diffère de moins de 20 ms. Les segments de silence ou de bruit sont ignorés. La valeur de ce critère est généralement plus grande pour les entrées correctes que pour celles incorrectes.

(a) entrée correcte : *… he will sleep …* (b) entrée incorrecte : *… he'll come …*

Figure 3. *Phonèmes pris en compte (indiqués par des flèches) pour calculer le pourcentage de segments phonétiques ayant la même étiquette dans l'alignement contraint (en haut) et dans l'alignement non contraint (en bas), pour un exemple d'entrée correcte (à gauche) et d'entrée incorrecte (à droite)*

3.1.2. *Critère associé aux trames*

Ce critère, illustré par la figure 4, mesure l'adéquation des alignements contraints et non contraints, non pas au niveau des segments, mais au niveau des trames. De plus, la comparaison est effectuée au niveau des classes phonétiques pour tenir compte du fait qu'il est assez fréquent que, lorsqu'un phonème est mal reconnu, il soit remplacé par un phonème de la même classe. Ce critère est égal au pourcentage de trames ayant leurs étiquettes appartenant à la même classe phonétique dans les deux alignements ; chaque classe phonétique correspondant à des sons qui partagent au moins une caractéristique phonétique comme par exemple les voyelles (V), les semi-voyelles (S), les fricatives (F), les nasales (N) et les occlusives (O). La valeur de ce critère est généralement plus grande pour les entrées correctes que pour celles incorrectes.

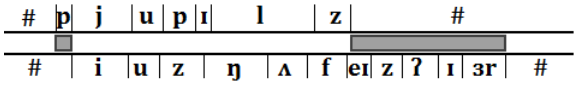
(a) entrée correcte : … *friends* … (b) entrée incorrecte : … *got* …

Figure 4. *Trames prises en compte (indiquées par des flèches) pour calculer le pourcentage de trames associées à la même classe phonétique dans les deux alignements, contraint (en haut) et non contraint (en bas), pour un exemple d'entrée correcte (à gauche) et d'entrée incorrecte (à droite)*

3.1.3. *Critère associé aux zones de non-parole*

Ce critère vise à mesurer les écarts entre les alignements au niveau des segments de non-parole. En effet, comme illustré par les zones grisées de la figure 5, pour une entrée incorrecte, l'alignement contraint doit souvent étirer les segments de silence ou au contraire les raccourcir pour s'accommoder des contraintes incorrectes (trop ou pas assez de mots dans la transcription par rapport à ce qui a été prononcé). Ce critère est égal à la différence de durée des segments de non-parole (silence ou bruit)

entre les deux alignements ; la différence est exprimée en pourcentage de la durée totale de l'énoncé. Elle est généralement plus petite pour les entrées correctes que pour celles incorrectes.



Entrée incorrecte : ... *pupils* ...

Figure 5. *Zones prises en compte (indiquées par des rectangles grisés) pour calculer la différence entre les segments de non-parole dans les deux alignements, contraint (en haut) et non contraint (en bas), pour un exemple d'entrée incorrecte*

3.1.4. *Critère du logarithme du rapport des vraisemblances*

Il exprime le fait que l'enchaînement des mots de l'alignement contraint est cohérent, ou non, avec le signal de parole. Si le texte ne correspond pas, l'alignement contraint conduit à une vraisemblance beaucoup plus petite que celle obtenue avec l'alignement non contraint. Ce critère est égal au logarithme du rapport des vraisemblances des deux alignements (contraint sur non contraint). Une valeur proche de zéro indique que les deux alignements conduisent à la même vraisemblance, ce qui signifie qu'ils correspondent à la même séquence de phonèmes et donc à une entrée correcte. La valeur de ce critère est plus petite (valeur négative) pour les entrées incorrectes.

3.1.5. *Critère associé aux phonèmes de durée minimale*

Ce critère, illustré par la figure 6, mesure les portions d'alignement extrêmes, c'est-à-dire celles qui ne peuvent pas être plus courtes à cause de la topologie même des modèles de Markov associés aux phonèmes (dans notre expérimentation, cela correspond à un minimum de trois trames pour un segment, minimum qui résulte des trois états émetteurs du modèle). Ce critère est égal à la différence entre le nombre de phonèmes ayant la durée minimale de trois trames dans les deux segmentations. La différence est exprimée en pourcentage du nombre total de phonèmes de l'alignement contraint. Une quantité importante de phonèmes avec des durées minimales pourrait indiquer des anomalies au sein de l'alignement. La valeur de ce critère est généralement plus petite pour les entrées correctes que pour celles incorrectes.

(a) entrée correcte : … *lambs* … (b) entrée incorrecte : … *pupils* …

Figure 6. *Phonèmes pris en compte (indiqués par des rectangles grisés) pour calculer la différence entre le nombre de phonèmes de durée minimale entre l'alignement contraint (en haut) et l'alignement non contraint (en bas), pour un exemple d'entrée correcte (à gauche) et d'entrée incorrecte (à droite)*

3.2 *Décision d'entrée correcte* vs *incorrecte*

Compte tenu des critères de comparaison choisis pour la décision et de la tâche de classification limitée à deux classes (correcte ou incorrecte), le modèle prédictif de la régression logistique (Dreiseitl et Ohno-Machado, 2002) a été choisi comme classifieur binaire. La régression logistique est utilisée ici pour estimer la probabilité d'une entrée d'être correcte connaissant les valeurs x_k des K critères de comparaison calculés pour l'énoncé courant :

$$f(\bar{X}) = \frac{1}{1 + \exp\left(-(\alpha_0 + \alpha_1 x_1 + \cdots + \alpha_K x_K)\right)} \quad [1]$$

Les paramètres du classifieur sont estimés sur les données d'apprentissage $D = \{\overline{X_n}, y_n\}, n = 1, \ldots, N$ où $\overline{X_n} = < x_1^n, x_2^n, \ldots, x_K^n >$ est le vecteur des K critères de comparaison calculés pour l'entrée à classifier, y_n indique l'appartenance de l'entrée à la classe correcte ($y_n = 1$) ou à la classe incorrecte ($y_n = 0$) et N précise le nombre d'entrées dans le corpus d'apprentissage.

Les paramètres $\alpha =< \alpha_0, \alpha_1, \ldots, \alpha_K >$ de la fonction de régression logistique sont déterminés en minimisant une fonction d'erreur E qui indique la discordance entre l'appartenance à une classe (correcte, 1, ou incorrecte, 0) et la valeur $f(X)$ de la fonction logistique, qui varie entre 0 et 1.

$$E = -\sum_{n=1}^{N} \left\{ y_n \, ln\left(f(\overline{X_n})\right) + (1 - y_n) \, ln\left(1 - f(\overline{X_n})\right) \right\} \quad [2]$$

La minimisation est effectuée par la méthode de la descente du gradient. Cet algorithme d'optimisation numérique vise à obtenir un optimum (éventuellement local) par améliorations successives. À partir d'un point de départ α_0, les paramètres sont modifiés itérativement jusqu'à atteindre la condition d'arrêt (amélioration de la fonction d'erreur plus petite qu'un seuil donné, ou nombre maximal d'itérations).

3.3 *Contexte expérimental*

L'approche proposée ci-dessus a été évaluée sur le corpus anglais de parole non native décrit dans la section 2. La partie apprentissage a servi pour l'estimation des paramètres α, et les évaluations de performances ont été menées sur la partie test.

Le corpus utilisé ne contient que des entrées correctes (même si la parole non native est sujette à de nombreux défauts de prononciation). Aussi, pour simuler les entrées incorrectes associées à des prononciations non attendues, nous avons utilisé les mêmes signaux audio, mais nous avons associé à chacun d'eux un texte qui ne lui correspond pas. Nous avons modifié les transcriptions de deux manières différentes : soit en remplaçant un mot ou une suite de mots (avec un choix de taille minimale de la séquence de trois, quatre, cinq, six ou sept syllabes), soit en remplaçant la phrase entière. Ces remplacements entre les mots et entre les phrases sont aléatoires.

Les outils HTK (Young *et al*., 2002) ont été utilisés pour décoder les signaux audio. Les vecteurs acoustiques des trames sont constitués de douze coefficients cepstraux (MFCC – *Mel Frequency Cepstrum Coefficients*) plus le logarithme de l'énergie, ainsi que les dérivées premières et secondes. L'analyse acoustique est effectuée sur des fenêtres de 32 ms, avec 10 ms de décalage entre trames. L'alignement texte-parole est obtenu avec une modélisation markovienne des phonèmes (une quarantaine d'unités anglaises, plus silence, bruit, et trois unités françaises : schwa, /ɑ̃/ et /y/) apprise sur les corpus TIMIT et ESTER2 (cf. section 2). Chaque état (densité) d'un modèle de Markov a été modélisé par un mélange de seize gaussiennes.

Deux lexiques de prononciations ont été utilisés. Le premier lexique, le lexique natif, inclut uniquement les variantes de prononciation natives (CMU DICTIONARY[8]). Le second lexique, le lexique non natif, inclut en plus les variantes de prononciation non natives observées dans la partie apprentissage du corpus de parole non native (Mesbahi *et al*., 2011).

L'analyse des performances de détection des entrées incorrectes est effectuée à partir des courbes DET (*Detection Error Tradeoff* – Martin *et al., 1997*) qui représentent l'évolution des taux de fausses acceptations et de rejets à tort (ou faux rejets) en fonction de différentes valeurs du seuil de décision. Le taux de fausses acceptations (FA) est le pourcentage d'entrées incorrectes classées « correctes » par le système. Le taux de faux rejets (FR) est le pourcentage d'entrées correctes classées « incorrectes » par le système. Une entrée est acceptée, *i.e.* classée « correcte », seulement si la valeur de la fonction de régression logistique $f(X)$ est supérieure à un seuil σ. Le choix de différentes valeurs du seuil $\sigma \in [0,1]$ permet de tracer les courbes DET. Finalement, par analogie avec la F-mesure (Chinchor,

[8] www.speech.cs.cmu.edu/cgi-bin/cmudict

1992), on calcule pour chaque courbe la valeur F qui maximise la moyenne harmonique des taux de bons rejets des entrées incorrectes $(1 - FA)$ et de bonne acception des entrées correctes $(1 - FR)$:

$$\frac{1}{F} = \frac{1}{2}\left(\frac{1}{1-FA} + \frac{1}{1-FR}\right) \quad [3]$$

3.4 *Évaluations et discussions*

Étant donné qu'il n'est pas possible de savoir à l'avance si la phrase prononcée sera entièrement, ou seulement partiellement, différente de la phrase attendue, nous avons optimisé les paramètres de la fonction logistique à partir d'exemples correspondant à des incohérences plus ou moins importantes entre le signal de parole et le texte (de trois syllabes incorrectes à toute la phrase incorrecte). Cela fournit une seule fonction de décision. Une analyse détaillée du comportement est ensuite effectuée sur chaque groupe séparément.

La figure 7 présente l'impact de la prise en compte des variantes de prononciation non natives dans le lexique des prononciations. Les évaluations des performances de rejet portent sur le sous-ensemble des entrées entièrement modifiées, et la décision exploite la combinaison de tous les critères de comparaison disponibles (voir détails plus loin). Les courbes montrent que l'utilisation de variantes de prononciation non natives dans le lexique améliore notablement la détection des entrées incorrectes.

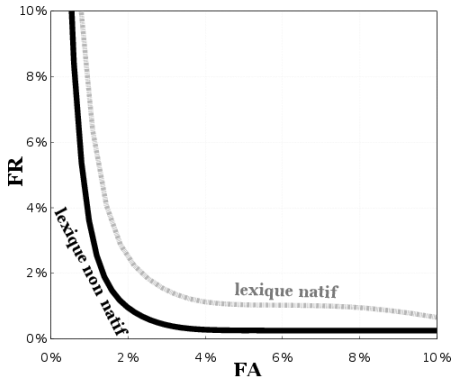


Figure 7. *Impact des variantes de prononciation non natives dans le lexique sur le rejet des entrées incorrectes (évaluation sur le sous-corpus des phrases entièrement incorrectes, et utilisation de tous les critères de comparaison)*

La figure 8 présente l'impact de la combinaison de différents critères de comparaison pour la détection des entrées incorrectes. Les résultats obtenus avec le critère standard correspondant au rapport des vraisemblances entre alignement

contraint et alignement non contraint servent de référence. La courbe « 8 critères » résulte de l'utilisation des quatre premiers critères décrits dans la section 3.1 (phonèmes, trames, non-parole et rapport des vraisemblances) calculés pour les deux variantes d'alignements non contraints (boucle de mots et boucle de phonèmes). La courbe « 10 critères » inclut en plus l'information sur les segments de durée minimale. Les expériences montrent que l'utilisation de tous les critères conduit aux meilleures performances de détection.

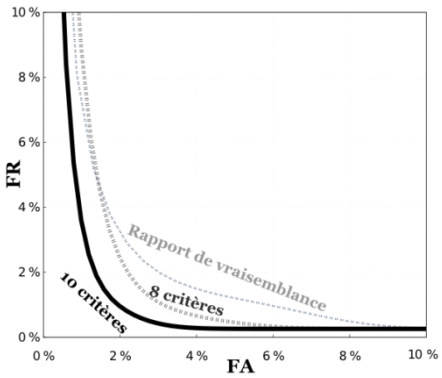


Figure 8. *Impact de la combinaison de différents critères de comparaison (évaluation sur le sous-corpus des phrases entièrement incorrectes, et utilisation du lexique avec prononciations non natives)*

La figure 9 présente une synthèse des performances globales de la détection des entrées incorrectes, en utilisant les prononciations non natives et les dix critères, en fonction de l'incohérence de l'entrée. L'incohérence correspond au nombre minimal de changements nécessaires (distance d'édition) pour mettre en correspondance la suite des phonèmes de la phrase prononcée avec la suite des phonèmes de la phrase attendue (les changements possibles sont l'insertion, la suppression et la substitution de phonèmes). Les résultats de détection sont regroupés en fonction de l'incohérence des entrées (plusieurs intervalles de distances d'édition sont considérés) et le graphique affiche les valeurs de F correspondantes. À partir d'une incohérence de plus de six changements de phonèmes, nous pouvons obtenir une performance supérieure à 80 %.

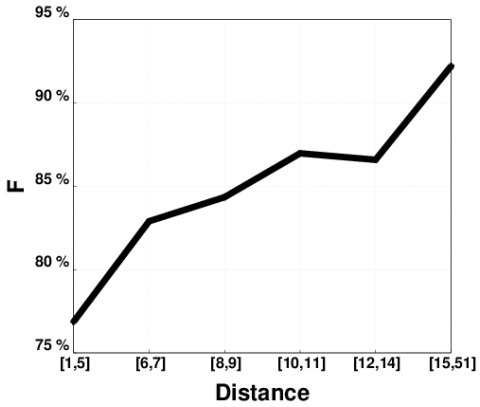


Figure 9. *Performances de détection des entrées incorrectes en fonction de l'incohérence (distance d'édition entre la suite des phonèmes de la phrase prononcée et celle de la phrase attendue*

En conclusion, les résultats de notre approche sur la comparaison entre les alignements non contraints et l'alignement contraint montrent que la combinaison de plusieurs critères de comparaison permet de faire beaucoup mieux que le simple rapport des vraisemblances pour le rejet des entrées incorrectes. L'approche tire bénéfice de l'introduction de variantes de prononciation non natives dans le lexique des prononciations. Cependant, elle ne permet pas un rejet efficace si l'incohérence entre le signal acquis et la phrase attendue porte sur un court segment de parole (typiquement moins de six phonèmes).

4. Alignement phonétique robuste aux déviations de la parole non native

Lorsque la phrase prononcée par l'apprenant a franchi l'étape de détection des entrées incorrectes, elle doit être segmentée phonétiquement afin de calculer, entre autres, des durées. Cette segmentation phonétique automatique qui repose sur un alignement texte-parole rencontre deux principales sources d'erreurs. La première tient à la tâche de segmentation en elle-même, la deuxième vient des multiples variantes de prononciation qui doivent être prises en considération pour traiter la parole non native. L'objectif de cette étude est d'obtenir un système d'alignement phonétique automatique qui fournisse des frontières temporelles précises tout en étant tolérant aux déviations des prononciations non natives de l'apprenant.

Lors de la prononciation d'une langue seconde, les erreurs (déviations) de prononciation viennent souvent de la confrontation de deux systèmes phonologiques, celui de la langue maternelle de l'apprenant (L1) et celui de la langue seconde (L2). Selon la théorie de Flege (1995), les phonèmes les plus difficiles à acquérir sont ceux qui n'existent pas dans le système phonologique de

l'apprenant (celui de sa langue maternelle L1) et qui sont proches d'autres phonèmes de ce système. Ainsi, les Français tendront à réaliser un « th » anglais (/ð/, /θ/) comme un /z/ou un /s/, mais auront moins de problème à réaliser un /h/ anglais, qui n'existe pas en français et ne se rapproche pas d'un son français. Une autre source d'erreur provient de la graphie. Parmi les très nombreux exemples d'erreurs provenant de la graphie, prenons le cas du /h/. Bien qu'ils soient capables de le réaliser quand ils l'entendent, les locuteurs français tendent à ne pas le prononcer, puisqu'en français le symbole correspondant au son s'écrit mais ne se prononce pas.

On retrouve bien ce type de déviations dans le corpus non natif, comme le montrent les quelques exemples du tableau 2. Les prononciations natives proviennent du CMU DICTIONARY ; elles correspondent à la prononciation d'un locuteur natif anglo-américain. Les prononciations non natives correspondent à des variantes fréquemment observées sur la partie apprentissage du corpus.

Mot	Prononciations natives	Prononciations non natives
mother	m ʌ ð ɚ	m o z œ ʁ
with	w i ð	w i z
lamb	l æ m	l ã b
has	h æ z	æ z

Tableau 2. *Exemple de prononciations natives et de variantes non natives*

4.1 *Système d'alignement phonétique et outil d'évaluation*

L'alignement phonétique résulte d'un alignement parole-texte qui repose sur des modèles acoustiques des phonèmes et un lexique de prononciations. Comme dans la section 3, les modèles acoustiques sont fondés sur des modèles de Markov cachés à trois états et exploitent le même paramétrage acoustique. Nous avons choisi d'utiliser des modèles acoustiques indépendants du contexte car diverses expériences, dont celles de Toledano *et al.* (2003), ont montré qu'ils permettaient d'obtenir une segmentation plus précise que celle résultant d'un alignement avec des modèles contextuels.

Nous avons évalué plusieurs modèles acoustiques de phonèmes correspondant à différents apprentissages des paramètres. Certains modèles ont été appris uniquement sur de la parole native, et d'autres adaptés sur la partie apprentissage du corpus de parole non native (cf. section 2). La méthode d'adaptation est fondée sur l'algorithme MLLR (*Maximum Likelihood Linear Regression* – Leggetter et Woodland, 1995). De plus, nous avons fait varier le nombre de gaussiennes par état de 1 à 8.

Nous avons également évalué plusieurs lexiques de prononciations : le lexique natif comprenant uniquement les variantes de prononciation natives (provenant du CMU DICTIONARY), et dix lexiques comprenant, en plus, les prononciations non natives observées au moins n fois sur la partie apprentissage du corpus de parole non native (n variant de 1 à 10). La figure 10 montre l'évolution du nombre moyen de variantes de prononciation par mot (incluant la ou les variantes natives) en fonction du nombre minimal d'observations (n) pour la sélection des variantes non natives. Ce nombre moyen de variantes par mot varie de 1,3 pour le lexique natif, jusqu'à 2,2 lorsque l'on considère les variantes observées au moins deux fois dans le corpus d'apprentissage et 3,3 si l'on considère toutes les variantes observées.

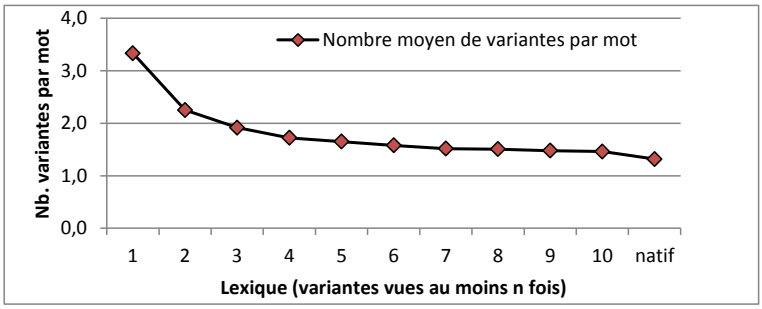


Figure 10. *Nombre moyen de variantes de prononciation par mot dans les lexiques*

La précision des différentes segmentations phonétiques de la partie test du corpus de parole non native a été évaluée avec l'outil CoALT (*Comparing Automatic Labelling Tool* – Fohr et Mella, 2012). CoALT a été développé pour comparer deux segmentations phonétiques automatiques à une segmentation phonétique de référence, et offre l'avantage de fournir des statistiques sur les résultats d'alignement et de mettre en évidence les différences dans les erreurs de décodage ou d'alignement. Cette analyse fine des erreurs de reconnaissance phonétique et de l'alignement est précieuse dans le cadre de la mise au point d'outils pour l'apprentissage d'une langue seconde. Une autre caractéristique importante de CoALT est que l'utilisateur définit ses propres critères d'évaluation et de comparaison des segmentations phonétiques automatiques par l'intermédiaire de règles. Il peut, par exemple, indiquer s'il privilégie la détection du bon phonème ou de la bonne classe de phonèmes ou bien s'il privilégie plutôt la précision temporelle des frontières des phonèmes. L'utilisateur peut également définir des classes d'équivalence lui permettant d'analyser uniquement certains cas de substitution, d'omission, ou d'insertion phonétique ou certaines frontières temporelles.

4.2 *Apport des variantes de prononciation non natives*

Nous avons étudié l'influence des variantes de prononciation non natives et de la méthode d'apprentissage, d'une part, sur le taux d'erreur des étiquettes phonétiques

(insertion, omission, confusion) et, d'autre part, sur la précision des frontières temporelles (pourcentage de frontières situées à plus de 20 ms de la frontière manuelle). Les classes phonétiques décrites dans le tableau 3 sont utilisées pour l'étude de la précision des frontières ; *i.e.* une frontière peut être considérée comme bien positionnée même si le son n'est pas correctement identifié tant que la substitution porte sur un son de la même classe.

Classe		Contenu
VOY	voyelles	ɑ æ ʌ ɔ ə e ɛ ɜ ɝ ɪ i ʊ u y ɑ̃ aʊ aɪ eɪ oʊ oɪ
SEM	semi-voyelles	w j
LIQ	liquides	l r
NAS	nasales	m n ŋ
OCC	occlusives	p t k b d g
AFF	affriquées	tʃ dʒ
FRI	fricatives	θ f s ʃ ð v z ʒ h
SIL	silence et bruit	(en fait toutes zones de non-parole)

Tableau 3. *Classes pour l'analyse de la précision de la segmentation phonétique*

La figure 11 synthétise une partie des résultats en présentant les performances obtenues avec des modèles acoustiques natifs et avec des modèles acoustiques adaptés sur des données non natives, pour différents lexiques incluant les variantes de prononciation vues au moins n fois (n variant de 1 à 10) dans la partie apprentissage du corpus, ou ne comportant que des variantes de prononciation natives.

Nous pouvons remarquer sur la partie gauche de la figure 11 que l'ajout de variantes non natives de prononciation a une forte influence sur le taux d'erreur des phonèmes (phonèmes résultant de l'alignement contraint texte-parole, avec prise en compte de variantes non natives dans le lexique des prononciations) alors que l'adaptation sur le corpus non natif a très peu d'influence. Les meilleurs résultats sont obtenus en prenant en compte les variantes observées au moins deux ou trois fois dans le corpus d'apprentissage non natif ; cela permet d'éviter les variantes aberrantes ou exceptionnelles qui dégradent les résultats.

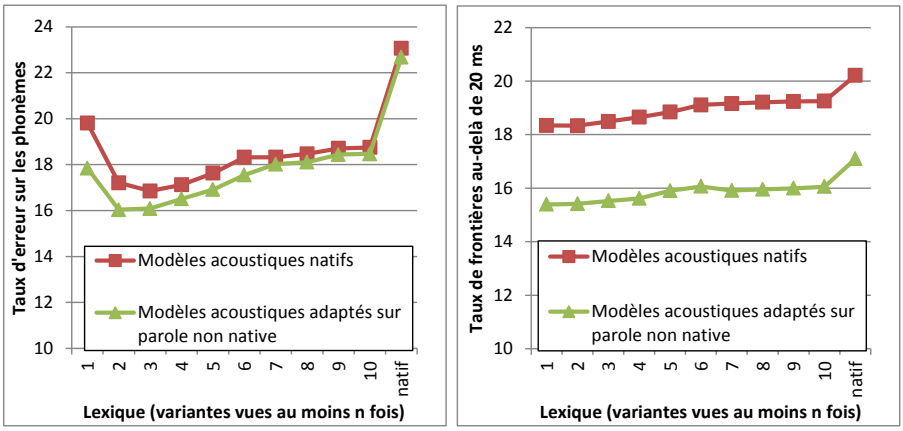


Figure 11. *Évaluation de la segmentation phonétique en fonction des variantes du lexique – variantes vues au moins n fois (n = 1...10) sur le corpus d'apprentissage, ou variantes natives uniquement*

En revanche, inclure des variantes de prononciation non natives a un faible impact sur la précision des frontières temporelles (cf. partie droite de la figure 11), alors qu'adapter les modèles acoustiques sur la parole non native améliore significativement cette précision quel que soit le lexique utilisé.

Cette étude préliminaire, incluant également d'autres tests non décrits ici, a permis de sélectionner le système de segmentation phonétique possédant les meilleures performances : modèles acoustiques à deux gaussiennes par état appris sur TIMIT et adaptés sur le corpus non natif avec quatre classes de régression, et lexique de prononciations incluant les variantes de prononciation apparaissant au moins deux fois dans le corpus non natif.

Avant d'analyser plus finement la précision des frontières temporelles en fonction des classes de phonèmes, dans le but de produire ultérieurement un retour fiable sur la prosodie, il est important de vérifier que le système ne commet pas trop d'erreurs de confusion entre les classes phonétiques. Pour cela nous avons construit la matrice de confusion présentée dans le tableau 4 en fonction des classes de phonèmes décrites dans le tableau 3. Celle-ci montre que la segmentation automatique fait très peu d'erreurs de confusion interclasse. Toutes les classes sauf les semi-voyelles et la classe SIL ont un taux d'erreur inférieur à 6 %.

	VOY	SEM	LIQ	NAS	OCC	AFF	FRI	SIL	*Omi.*	Nb. occ.
VOY	**97,1**	0,1	0,0	0,3	0,0	0,1	0,1	1,0	*1,3*	4 301
SEM	2,6	**82,3**	10,9	0,0	0,0	0,0	1,9	0,0	*2,3*	266
LIQ	0,6	0,2	**97,1**	0,1	0,2	0,0	0,6	0,7	*1,1*	852
NAS	0,0	0,0	0,2	**98,2**	0,4	0,0	0,0	0,1	*1,1*	1 131
OCC	0,0	0,1	0,0	0,1	**96,8**	0,1	0,6	0,2	*2,1*	1 698
AFF	0,0	0,0	0,0	0,0	1,9	**94,4**	2,8	0,0	*0,9*	106
FRI	0,1	0,1	0,1	0,1	0,4	0,1	**97,2**	0,1	*1,9*	1 629
SIL	0,5	0,1	0,1	0,1	1,1	0,0	4,5	**70,8**	*22,8*	1 512
Ins.	*14,6*	*4,3*	*12,9*	*3,1*	*11,5*	*0*	*21,0*	*32,7*		*419*

Tableau 4. *Matrice de confusion (en %) entre classes (voir définition des classes dans tableau 3). Les insertions sont indiquées en pourcentage du nombre total d'insertions (419)*

5. *Fiabilité des frontières phonétiques selon les classes de sons*

La segmentation entre deux sons pose des problèmes liés à la nature des éléments en présence. En effet, la segmentation peut être relativement précise entre deux sons qui ont des modes d'articulation très différents, et, par voie de conséquence, des structures acoustiques très distinctes dont le passage se caractérise par une rupture abrupte (cf. partie gauche de la figure 12, qui représente une suite voyelle-occlusive-voyelle). En revanche, la segmentation ne peut pas être précise entre deux sons de mode d'articulation identique ou proche qui sont caractérisés par des structures acoustiques semblables et pour lesquels la transition d'une articulation à une autre, et d'une structure acoustique à une autre se fait de manière continue (cf. partie droite de la figure 12 qui représente une suite de quatre sons : voyelle-consonne liquide-voyelle-semi-voyelle).

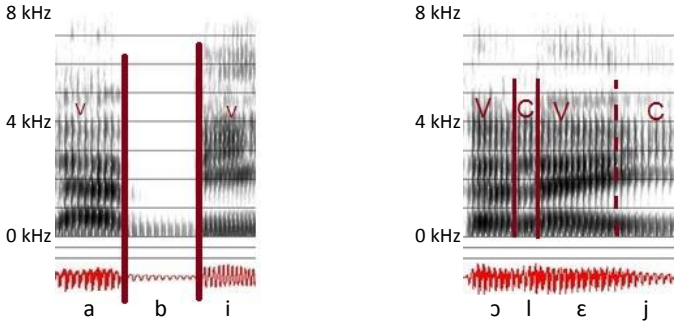


Figure 12. *Segmentation phonétique évidente (ex.* /abi/*) ou ambiguë (ex.* /ɔlɛj/*) ; l'extrait de gauche correspond à 240 ms de signal, et l'extrait de droite à 275 ms*

Nous avons donc évalué la précision des frontières obtenues par notre système d'alignement choisi au paragraphe précédent. Le but de cette évaluation est de déterminer les paires de classes de sons pour lesquelles la frontière peut être estimée avec une précision suffisante pour calculer des durées fiables, afin de pouvoir fournir à l'apprenant un retour prosodique pertinent. Pour les paires de classes de sons présentant des frontières imprécises aucun retour prosodique ne sera envoyé à l'apprenant afin d'éviter des retours erronés qui sont très préjudiciables dans le cas de l'apprentissage automatique d'une langue seconde. Pour cette évaluation, nous avons calculé et analysé les écarts entre les frontières de sons (frontière de début et frontière de fin) manuelles et celles mises par notre système.

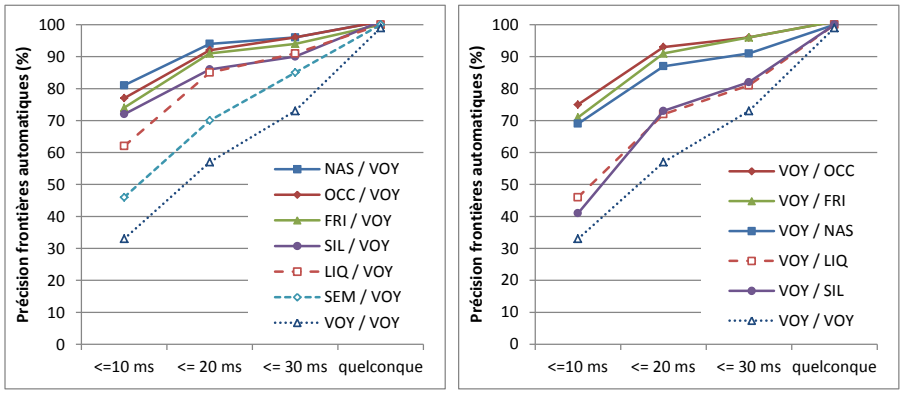


Figure 13. *Précision des frontières pour différentes paires de classes observées plus de cent fois dans le corpus d'évaluation (les courbes indiquent le pourcentage de frontières automatiques situées à moins de n ms de la position de référence)*

La figure 13 indique le pourcentage de frontières pour lesquelles cet écart est inférieur à un certain laps de temps (10, 20, 30 ms et plus), pour des séquences de sons avec une voyelle en contexte droit (*-VOY, partie gauche de la figure), et une voyelle en contexte gauche (VOY-*, partie droite de la figure) ; en se limitant aux séquences de classes observées au minimum cent fois dans le corpus d'évaluation.

La figure 14 montre l'histogramme des frontières placées avec un écart de moins de 20 ms par rapport à la frontière de référence pour différentes séquences de sons observées plus de cent fois dans le corpus d'évaluation. De nombreuses publications, dont (Hosom, 2009), indiquent un taux de plus de 90 % d'écart inter-annotateur inférieur à 20 ms pour les frontières des sons.

Comme nous pouvons le constater sur ces figures, les meilleures segmentations correspondent aux séquences dont les sons ont des modes d'articulation très différents : les séquences composées d'une consonne nasale, fricative ou occlusive suivie ou précédée d'une voyelle sont les mieux segmentées ; les écarts les plus importants sont obtenus pour les suites composées de deux voyelles, d'une semi-

voyelle et d'une voyelle, ou d'une occlusive et d'une fricative. Les résultats sont donc conformes à notre attente.

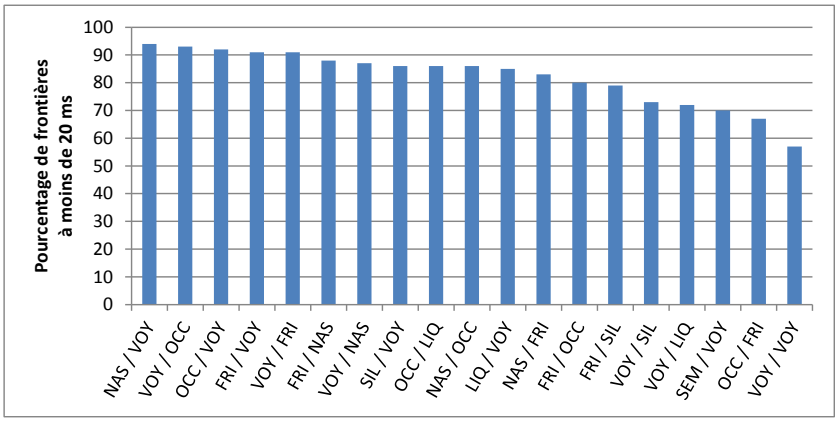


Figure 14. *Précision des frontières (pourcentage de frontières automatiques à moins de 20 ms de la frontière de référence) pour les paires de classes observées plus de cent fois dans le corpus d'évaluation*

Cette conclusion concerne le système de segmentation correspondant à la meilleure configuration. Afin d'établir des conclusions plus générales, moins dépendantes d'un système de segmentation, une autre étude a consisté à analyser la précision des frontières entre classes de sons pour un grand nombre d'étiqueteurs automatiques obtenus en faisant varier l'apprentissage ou l'adaptation des modèles acoustiques, le nombre de gaussiennes et les variantes non natives de prononciation. Au total quatre cents systèmes différents ont été évalués. Pour chaque étiqueteur, le logiciel CoALT a calculé le pourcentage de frontières automatiques situées à moins de 10 ms et à moins de 20 ms de la frontière de référence. Le tableau 5 présente la valeur médiane de ce pourcentage sur tous les étiqueteurs pour les frontières occlusive-voyelle et fricative-voyelle.

		Nombre total d'occurrences	Pourcentage ayant un écart	
			<= 10 ms	<= 20 ms
Occlusives	sourdes	400	78	93
	voisées	230	70	88
Fricatives	sourdes	300	82	94
	voisées	330	69	91

Tableau 5. *Médiane, calculée sur tous les étiqueteurs, du pourcentage de frontières occlusive-voyelle et fricative-voyelle situées à moins de 10 ou 20 ms de la frontière de référence*

Nous pouvons observer que les frontières consonne-voyelle sont plus précises pour les consonnes sourdes ce qui permettra, dans un tel contexte, de faire un retour plus fiable à l'apprenant sur la durée de ses voyelles notamment dans le cas des langues possédant des voyelles brèves et longues. La meilleure précision obtenue pour les consonnes sourdes peut s'expliquer par le fait que le bruit d'explosion ou de friction est plus intense pour les consonnes sourdes.

6. Conclusion

Dans cet article, nous avons essayé de prendre en compte la gestion d'un certain nombre d'erreurs pouvant intervenir dans un système automatique d'apprentissage de la prosodie d'une langue étrangère.

La première erreur que le système doit savoir traiter est l'erreur due à un problème d'acquisition du signal ou à l'apprenant qui ne respecte pas les consignes et qui ne prononce pas la phrase attendue. Il est impératif de détecter ce type d'erreur pour ne pas faire un retour erroné à l'apprenant. L'approche choisie pour le rejet de ces entrées incorrectes est fondée sur la comparaison entre un alignement contraint (par la phrase attendue) et un alignement plus libre (résultant d'un décodage guidé par une boucle de phonèmes ou une boucle de mots). Les résultats montrent que la combinaison de plusieurs critères de comparaison permet de faire beaucoup mieux que le simple rapport des vraisemblances pour le rejet des entrées incorrectes. L'approche est robuste à la parole non native, mais elle ne permet pas un rejet efficace si l'incohérence entre le signal acquis et la phrase attendue porte sur une courte portion de parole de moins de six phonèmes.

Pour effectuer des retours pertinents concernant la prosodie de l'apprenant, il est nécessaire d'obtenir une segmentation phonétique précise, par exemple pour calculer des durées. Or, un apprenant prononce parfois une suite de phonèmes différente de celle de la prononciation standard pour un mot (ajout, omission ou substitution d'un ou de plusieurs phonèmes) ce qui va entraîner des erreurs lors de l'alignement automatique. Pour pallier ces erreurs d'alignement, nous avons choisi d'ajouter des variantes de prononciation non natives dans le lexique des prononciations. Les expériences montrent qu'il vaut mieux se limiter aux variantes suffisamment fréquentes, ce qui permet d'éviter des variantes de prononciation trop atypiques qui viennent pénaliser l'alignement phonétique de certaines phrases.

L'analyse des erreurs d'alignement (frontières de phonèmes situées à plus de 10 ou 20 ms de la frontière manuelle) a permis de déterminer les frontières qui sont les plus précises, et qui permettront de faire un retour fiable à l'apprenant. Les frontières consonne-voyelle et voyelle-consonne montrent une bonne précision pour les fricatives, les occlusives et les nasales. De plus, nous avons examiné la précision des frontières fricative-voyelle et occlusive-voyelle pour un grand nombre d'étiqueteurs automatiques. Les frontières avec les fricatives et occlusives sourdes se sont avérées plus précises que celles obtenues avec les sonores. Ces résultats

peuvent aider à définir des exercices pour lesquels les retours prosodiques fournis à l'apprenant seront plus pertinents. Ces résultats permettront aussi d'éviter au système de faire des retours qui reposent sur des frontières dont on sait qu'elles ne sont pas suffisamment fiables.

Plusieurs prolongements de ces travaux sont en cours ou prévus. Le premier porte sur la mise au point d'un système automatique de génération de variantes de prononciation non natives. Le second thème porte sur l'élaboration de mesures de confiance sur les frontières de la segmentation automatique. Il serait en effet très utile de disposer d'une telle mesure pour ajuster dynamiquement les diagnostics et le choix des retours pertinents, au lieu de s'appuyer sur une tendance générale sur la précision des frontières entre classes.

Remerciements

Les travaux présentés dans cet article ont bénéficié du support du projet ALLEGRO (www.allegro-project.eu) financé par le programme européen INTERREG IV.

7 Bibliographie

Bisani M., Ney H., « Open vocabulary speech recognition with flat hybrid models », *Proceedings INTERSPEECH'2005*, Lisbonne, Portugal, p. 725-728, 2005.

Boite R., *Traitement de la parole*, Presses polytechniques et universitaires romandes, Lausanne, 2000.

Bonneau A., Colotte V., « Automatic Feedback for L2 Prosody Learning », in *Speech Technologies, Book 2*, I. IPSIC (editor), Intech, p. 55-70, June 2011.

Bot K. (de), « Visual feedback on intonation I: Effectiveness and induced practice behaviour », *Language and Speech*, vol. 26, n° 4, p. 331-350, 1983.

Chinchor N., « MUC-4 evaluation metrics », in *Proc. of the Fourth Message Understanding Conference*, San-Francisco, CA, USA, p. 22-29, 1992.

Chun D. M., « Signal analysis software for teaching discourse intonation », *Language Learning and Technology*, vol. 2, n° 1, p. 61-77, 1998.

Dargnat M., Bonneau A., Colotte V., « Perception et apprentissage des contours prosodiques en l1 et en l2 », http://mathilde.dargnat.free.fr/INTONALE/intonale-web.html, 2010.

Davel M., van Heerden C., Barnard E., « Validating smartphone-collected speech corpora », *Proceedings Workshop on Spoken Languages Technologies, SLTU'2012*, Monkey Valley Resort, South Africa, p. 68-75, 2012.

Dreiseitl S., Ohno-Machado L., « Logistic regression and artificial neural network classification models », *Journal of Biomedical Informatics*, vol. 35, p. 352-359, 2002.

Eskenazi M., « An overview of spoken language technology for education », *Speech Communication*, vol. 51, 2009, p. 832-845.

Flege J.E., « Second Language Speech Learning: Theory, Findings and Proble ms. In: Strange », W. (ed). Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-Language Speech Research, p. 233-272. Timonium: York Press, 1995.

Fohr D., Mella O., « CoALT : A Software for Comparing Automatic Labelling Tools » *Proceedings of the Eight International Conference on Language Resources and Evaluation*, Istanbul, Turquie, 2012.

Galliano S., Gravier G., Chaubard L., « The ESTER 2 evaluation campaign for rich transcription of French broadcasts », *Proceedings INTERSPEECH'2009*, Brighton, UK, 2009.

Garofolo J., Lamel L., Fisher W., Fiscus J., Pallett D., Dahlgren N., Zue V., « TIMIT Acoustic-Phonetic Continuous Speech Corpus », LDC, Philadelphie, USA, 1993.

Germain-Rutherford A. Martin, P., « Présentation d'un logiciel de visualisation pour l'apprentissage de l'oral en langue seconde », *ALSIC, Apprentissage des langues et systèmes d'information et de communication*, vol. 3, n° 1, p. 71-86, 2000.

Henry G., Bonneau A., Colotte V., « Tools devoted to the acquisition of the prosody of a foreign language », *Proceedings International Congress of Phonetic Sciences ICPhS'2007*, Sarrebruck, Allemagne, p. 1593-1596, 2007.

Hosom J.-P., « Speaker-independent phoneme alignment using transition-dependent states », *Speech Communication*, vol. 51, p. 352-368, 2009.

James E., « The Acquisition of a Second-Language Intonation Using a Visualizer », *Canadian Modern Language Review*, vol. 33, n° 4, p. 503-506, 1977.

Lecouteux B., Linarès G., « Combined low level and high level features for OOV word detection », *Proceedings INTERSPEECH*, 2009, Brighton, UK, 2009.

Leggetter C.J., Woodland P.C., « Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models », *Computer Speech and Language*, vol.9, n° 2, p. 171-185, avril 1995.

Martin A., Doddington G., Kamm T., Ordowski M., Przybocki M. « The DET Curve in Assessment of Detection Task Performance », *Proceedings EUROSPEECH 1997*, Rhodes, Grèce, vol. 4, p. 1899-1903, sept. 1997.

Marty F., « Les enseignants de langues face à l'Enseignement Assisté par Ordinateur », *Le Journal de la Formation Continue et de l'EAO*, 169, déc. 1983.

Mesbahi L., Jouvet D., Bonneau A., Fohr D., Illina I., Laprie Y., « Reliability of non-native speech automatic segmentation for prosodic feedback », *Proceedings workshop on Speech and Language Technology in Education SlaTE 2011*, Venise, Italie, 2011.

Orosanu L., Jouvet D., Fohr D., Illina I., Bonneau A., « Détection de transcriptions incorrectes de parole non native dans le cadre de l'apprentissage de langues étrangères », *Actes JEP-TALN-RECITAL 2012*, Grenoble, France, juin 2012.

Pinto J., Szoke I., Prasanna S.R.M., Hermansky H., « Fast approximate spoken term detection from sequences of phonemes », *Proceedings ACM SIGIR Workshop on Searching Spontaneous Conversational Speech*, Singapour, p. 28-33, 2008.

Speech Communication, special issue on "Spoken language technology for education", vol. 51, 2009.

Toledano D., Gomez, L. Grande L., « Automatic phonetic segmentation », *IEEE Transaction on Speech and Audio Processing*, vol. 11, p. 617-625, 2003.

Vardanian R., « Teaching English Intonation through Oscilloscope Displays », *Language Learning*, vol. 14, n° 3-4, p. 109-117, 1964.

White C., Zweig G., Burget L., Schwarz P., Hermansky H., « Confidence estimation, OOV detection and language ID using phone-to-word transduction and phone-level alignments », *Proceedings ICASSP'2008*, Las Vegas, Nevada, USA, p. 4085-4088, 2008.

Young S., Evermann G., Kershaw D., Moore G., Odell J., Ollason D., Povey D., Valtchev V., Woodland P., « *The HTK Book (for HTK version 3.2)* », Cambridge University Engineering Department, 2002.

Zahorian S. A., Hu H., « A spectral/temporal robust method for robust fundamental frequency tracking », *Journal of Acoustical Society of America*, vol. 123, n° 6, p. 4559-4571, 2008.