Détection de transcriptions incorrectes de parole non-native dans le cadre de l'apprentissage de langues étrangères

Luiza Orosanu Denis Jouvet Dominique Fohr Irina Illina Anne Bonneau
INRIA - LORIA, 615 rue de Jardin Botanique 54600 Villers-les-Nancy
{luiza.orosanu, denis.jouvet, dominique.fohr, irina.illina,
anne.bonneau}@loria.fr

RÉSUMÉ

Cet article analyse la détection de transcriptions incorrectes de parole non-native dans le contexte de l'apprentissage de langues étrangères. L'objectif est de détecter et rejeter les transcriptions incorrectes (i.e. celles pour lesquelles le texte ne correspond pas au signal de parole associé) tout en étant tolérant aux défauts de prononciation inhérents à la parole non-native. L'approche proposée exploite la comparaison d'un alignement contraint par la transcription à vérifier avec un alignement non contraint correspondant à un décodage phonétique. Plusieurs critères de comparaison sont décrits et combinés par l'intermédiaire d'une fonction de régression logistique. L'article analyse l'influence de divers paramétrages comme l'impact des variantes de prononciation non-natives, l'utilisation de fonctions de décision spécifiques à la longueur des transcriptions, et l'impact d'un apprentissage de la fonction de décision avec la parole native ou non-native. Les évaluations de performances sont menées à la fois sur des corpus de parole natives et non-natives.

ABSTRACT

Detection of incorrect transcriptions of non-native speech in the context of foreign language learning

This article analyses the detection of incorrect transcriptions of non-native speech in the context of foreign language learning. The purpose is to detect and reject incorrect transcriptions (i.e. those for which the text does not correspond to the associated speech signal) while being tolerant to the pronunciation defects of non-native speech. The proposed approach exploits the comparison between two alignements : one constrained by the transcript which is being checked, with an other one unconstrained, corresponding to a phonetic decoding. Several criteria are described and combined via a logistic regression function. The article analyzes the influence of different settings, such as the impact of non-native pronunciation variants, the use of decision functions dependent on the length of the transcriptions, and the impact of learning decision functions on native or non-native speech. The performance evaluations are conducted both on native speech and non-native speech.

MOTS-CLÉS : Apprentissage d'une langue étrangère, entrées incorrectes, parole non-native, variantes de prononciation, alignements contraint et non-contraint.

KEYWORDS: Foreign language learning, incorrect transcriptions, non-native speech, pronunciation variants, constrained and unconstrained alignements.

1 Introduction

L'aide à l'apprentissage des langues étrangères est un domaine d'application de la reconnaissance automatique de la parole qui s'est développé ces dernières années. L'objectif est de détecter et signaler à l'apprenant ses erreurs ou défauts de prononciation, afin qu'il puisse les corriger, et peu à peu améliorer sa maîtrise de la langue étrangère. L'une des principales difficultés pour ces systèmes est la détection et la localisation automatique des défauts de prononciation (Herron *et al.*, 1999) tout en restant robuste à la parole non-native. Des méthodes ont été proposées pour déterminer un score de qualité de prononciation (Witt et Young, 2000) en exploitant des rapports de vraisemblance. De tels systèmes tirent profit de l'introduction de modèles acoustiques de phonèmes de la langue maternelle (en complément des modèles des phonèmes de la langue cible) ainsi que de la connaissance des défauts (variantes) possibles de prononciation non-natives.

Un autre élément important de l'apprentissage des langues concerne la prosodie. Certains projets ont porté sur le retour d'information sur les erreurs de durée (Eskenazi *et al.*, 2000) mais le retour d'information prosodique se résume fréquemment à jouer ou rejouer une prononciation du mot ou de la phrase par un locuteur natif. Une méthode originale a été proposée dans (Henry *et al.*, 2007) qui vise à améliorer simultanément la production et la perception en combinant un retour prosodique précis et détaillé et un retour sonore basé sur une modification prosodique de la prononciation de l'apprenant. Cette approche nécessitant une segmentation phonétique de la prononciation de l'apprenant, une étude de la pertinence de la segmentation phonétique a été entreprise (Mesbahi *et al.*, 2011). Ces méthodes automatiques de diagnostic des prononciations reposent sur une segmentation phonétique du signal de parole qui est obtenue par alignement forcé du signal de parole avec les modèles correspondant à la phrase prononcée. La prise en compte de variantes de prononciation non natives améliore la qualité des alignements (Jouvet *et al.*, 2011).

Cependant, il arrive que le signal acoustique ne corresponde pas à la phrase attendue (erreur de prononciation, parole parasites, problème de capture du son, ...). Le système doit donc être capable de déterminer si le signal audio correspond ou pas à la phrase attendue. Ce type de décision correspond typiquement au rejet des entrées incorrectes ou des mots hors vocabulaire en reconnaissance de la parole (Bazzi et Glass, 2000; Boite, 2000). Contrairement à ces approches, qui visent essentiellement la parole native, ici nous voulons offrir un soutien à l'apprentissage des langues étrangères, et donc nous avons besoin de détecter les incohérences (i.e. un signal audio ne correspondant pas à la phrase attendue), mais en même temps tolérer les défauts de prononciations non-natives.

Donc, l'objectif de cet article est d'étudier en détail le rejet de transcriptions incorrectes dans le contexte de l'apprentissage des langues étrangères. La première partie présente une description de la méthodologie mise en œuvre, et en particulier les critères utilisés et la fonction de décision choisie. La deuxième partie du papier est consacrée à la description des expériences menées et à la discussion des résultats.

2 Méthodologie

Afin de rejeter les transcriptions incorrectes, tout en acceptant celles qui sont correctes, il faut déterminer si le signal audio et la transcription correspondent. Pour cela, nous avons choisi de

décoder les signaux audio de deux façons différentes. Tout d'abord, nous effectuons un décodage contraint, où l'on force le système à suivre la séquence des mots présents dans la transcription de référence. Ensuite, nous effectuons un décodage non-contraint, où l'on donne au système la liberté de choisir n'importe quel phonème pour n'importe quelle position dans la phrase. Finalement, nous comparons les deux alignements (contraint et non-contraint) afin de décider si la transcription est correcte ou non (i.e. si elle correspond ou pas au signal audio).

2.1 Critères pour la décision

Cette partie décrit les critères choisis pour différencier les transcriptions correctes de celles qui sont incorrectes. Ces critères sont basés sur des informations provenant des alignements contraints et non-contraints, en considérant les phonèmes, les trames ou les zones annotées silence/bruit.

1. Critère associé aux phonèmes : pourcentage de segments phonétiques qui ont le même label dans les deux alignements et dont au moins une limite temporelle diffère de moins de 20 ms. Les segments de silence/bruit sont ignorés. Sa valeur est bien plus grande pour les transcriptions correctes que pour celles incorrectes.

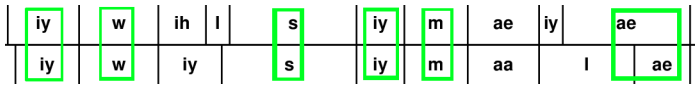

FIGURE 1 – Exemple de transcription correcte avec ses deux décodages : contraint (en haut) et non-contraint (en bas). Les rectangles verts indiquent les phonèmes pris en compte pour calculer le «critère associé aux phonèmes »

2. Critère associé aux trames : basé sur l'étiquetage des trames. Même si le décodage phonétique ne trouve pas le bon phonème, il est susceptible de le remplacer avec un phonème de la même classe. Une classe phonétique est représentée par des sons qui partagent au moins une caractéristique phonétique, et en particulier le «mode d'articulation". Nous calculons alors le pourcentage de trames ayant leurs étiquettes appartenant à la même classe phonétique. Ce pourcentage est donc généralement plus grand pour les transcriptions correctes que pour celles incorrectes. Dans l'exemple suivant nous avons considéré les classes phonétiques : voyelles (V), semi-voyelles (SV), fricatives (F), nasales (N) et plosives (P).

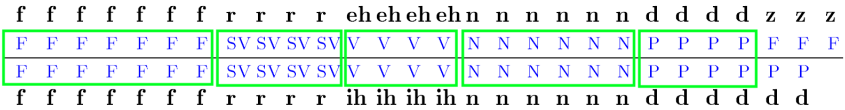

FIGURE 2 – Exemple de transcription correcte avec ses deux décodages : contraint (en haut) et non-contraint (en bas). Les rectangles verts indiquent les trames prises en compte pour calculer le «critère associé aux trames »

3. Critère associé aux zones de non-parole : basé uniquement sur les segments de non-parole (silence / bruit). Lorsque l'on force le système à aligner un signal audio sur un texte qui ne

lui correspond pas (le cas d'une transcription incorrecte), il est fréquent que le système ajoute plusieurs segments de silence entre les mots et/ou qu'il augmente ou diminue la durée de ceux qui existent réellement. Nous calculons donc la différence de recouvrement des segments de non-parole entre les deux alignements (exprimée en pourcentage du nombre total de trames). La valeur de ce critère sera plus petite pour les transcriptions correctes que pour celles incorrectes.

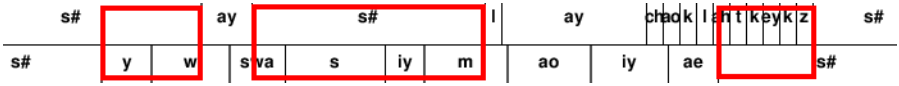

FIGURE 3 – Exemple de transcription incorrecte avec ses deux décodages : contraint (en haut) et non-contraint (en bas). Les rectangles rouges indiquent les trames prises en compte pour calculer le «critère associé aux zones de non-parole »

2.2 Classification de données

Compte tenu des critères choisis pour la décision (section 2.1) 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. En général, la régression logistique est utilisée pour calculer la probabilité d'appartenance à une classe parmi deux :

$$P\left(1|\overline{X},\alpha\right) = f(\overline{X}) = \frac{1}{1+exp(-(\alpha_0+\alpha_1 x_1+\alpha_2 x_2+\alpha_3 x_3))}$$

La première étape de l'approche consiste à apprendre les paramètres du classifieur sur les données d'apprentissage. Le corpus d'apprentissage est représenté par l'ensemble de données $D = \left\{\overline{X_i},\ y_i\right\},\ i = 1,...,N$ où :

– $\overline{X} =< x_1, x_2, x_3 >$ est le vecteur comprenant les informations sur la transcription à classifier, c'est-à-dire les critères par phonèmes, par trames et par zones de non-parole
– y indique l'appartenance à la classe correcte ($y = 1$) ou à la classe incorrecte ($y = 0$)
– N est le nombre de transcriptions dans le corpus d'apprentissage.

Les paramètres $\overline{\alpha} =< \alpha_0, \alpha_1, \alpha_2, \alpha_3 >$ sont déterminés par l'estimation du maximum de vraisemblance, qui se calcule en minimisant la fonction d'erreur :

$$E = -\sum_{i=1}^{N}\left(y_i \cdot ln\left(f(\overline{X_i})\right) + \left(1-y_i\right) \cdot ln\left(1-f(\overline{X_i})\right)\right)$$

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. A partir d'un point de départ α et une valeur initiale du pas de descente, les paramètres sont modifiés jusqu'à atteindre la condition d'arrêt (plus d'amélioration possible).

Apres, les courbes DET («detection error tradeoff ») sont utilisées pour présenter les résultats obtenus sur les données de test . Une courbe DET est un graphique des taux d'erreur pour les systèmes de classification binaire. Elle est utilisée ici comme un moyen de représenter les performances sur la tâche de classification des transcriptions, qui implique un compromis entre les taux de «*fausse acceptation* » (*FA*, le pourcentage de transcriptions incorrectes, mais classées

comme étant correctes par le système) et de «*faux rejet* » (*FR*, le pourcentage de transcriptions correctes, mais classées comme étant incorrectes par le système). Une transcription est acceptée seulement si la valeur de la régression logistique $f(\overline{X})$ est supérieure à un seuil σ . Pour tracer la courbe DET, différentes valeurs du seuil $\sigma \epsilon [0,1]$ sont utilisées. Les taux d'erreurs (FA, FR) pour chaque valeur du seuil sont indiqués sur le graphique. Finalement, le meilleur compromis entre les taux d'erreur (parmi tous les points disponibles sur les graphiques *DET*), est celui qui maximise la F-mesure :

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

3 Expériences et résultats

3.1 Contexte expérimental

Afin d'évaluer les approches mentionnées dans la section 2, deux corpus anglais de parole native et non-native sont utilisés. Ils proviennent du projet INTONALE (Dargnat *et al.*, 2010) consacré à l'étude prosodique. Le corpus natif contient environ 1500 signaux audio qui ont été enregistrés par 22 locuteurs (15 femmes et 7 hommes) anglais (66 phrases par locuteur). Le corpus non-natif contient environ 800 signaux audio qui ont été enregistrés par 34 locuteurs (29 femmes et 5 hommes) français (23 phrases par locuteur). Ces enregistrements ont été faits dans une pièce calme. Le logiciel développé pour l'enregistrement des mots ou des phrases affiche sur l'écran la phrase à prononcer. Le locuteur peut ensuite choisir, après chaque prononciation d'une phrase, de la répéter (en cas de problème) ou de passer à la suivante.

Les corpus utilisés contiennent tous des transcriptions correctes (même si la parole non-native est sujette à beaucoup de défauts de prononciation). Pour simuler des transcriptions incorrectes, nous utilisons les mêmes signaux audio, mais nous attachons à chacun une transcription qui ne lui correspond pas (tirée de façon aléatoire parmi les autres). Nous avons donc la même quantité de données correctes et incorrectes.

Chaque corpus, natif ou non-natif, est découpé en deux parties égales : une partie pour faire l'apprentissage des paramètres $\overline{\alpha}$, et l'autre pour évaluer les performances. Afin d'étudier la dépendance des paramètres à la longueur des transcriptions (nombre des phonèmes), chaque corpus est de nouveau découpé en 3 sous-ensembles : transcriptions courtes (moins de 19 phonèmes), transcriptions moyennes et transcriptions longues (plus de 30 phonèmes).

Les outils HTK (Young *et al.*, 2002) sont utilisés pour le décodage des signaux audio. Les modèles acoustiques ont été appris en utilisant le corpus anglais TIMIT (Garofolo, 2000). Ses signaux audio ont été enregistrés par 630 locuteurs américains, avec une fréquence d'échantillonnage de 16 kHz. L'analyse acoustique MFCC (Mel Frequency Cepstral Coefficients) donne 12 paramètres MFCC et le logarithme de l'énergie par trame (fenêtre de 32 ms, décalage de 10ms). La segmentation phonétique d'une transcription (décodage contraint) est obtenue avec des modèles acoustiques HMM (Hidden Markov Models) et la prise en compte des variantes de prononciation de chaque mot. Chaque état d'un modèle HMM a été modélisé par un mélange de 16 gaussiennes.

Deux lexiques ont été utilisés : le premier inclut seulement les variantes natives pour chaque mot (lexique natif : *CMU dictionary* (Hunt, 1996)) et le second inclut en plus des variantes non-natives (lexique non-natif). Un grand nombre de variantes de prononciation non-natives

observées dans le corpus de parole non-native ont été incluses dans le lexique de prononciation (Mesbahi *et al.*, 2011) (la génération automatique de variantes de prononciation non-natives sera étudiée dans les travaux futurs).

3.2 Évaluations

Cette partie étudie l'impact de différents paramètres de l'approche, en particulier pour le traitement de la parole non-native : l'impact du lexique de prononciation natif ou avec variantes non-natives, l'impact d'une fonction de décision globale, ou d'une fonction dépendante de la longueur de la transcription traitée et l'impact du type des données (natives ou non-natives) utilisées pour l'apprentissage des paramètres.

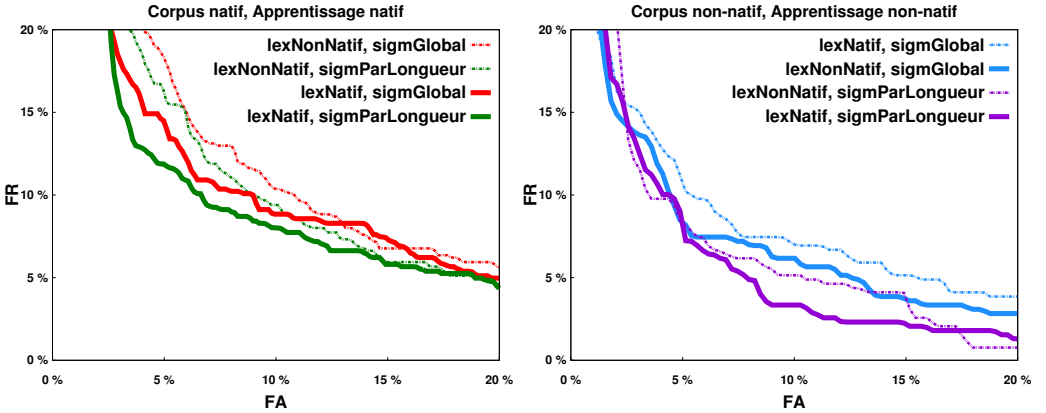

FIGURE 4 – Courbes DET pour les lexiques natif et non natif, et le paramétrage global (courbes rouges / bleues) ou dépendant de la longueur des transcriptions (courbes vertes / mauves)

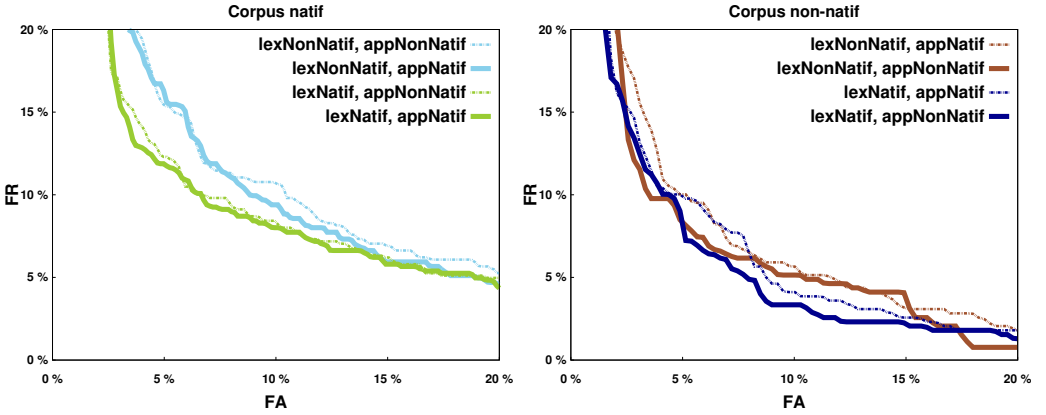

FIGURE 5 – Courbes DET pour l'apprentissage sur les corpus natif et non-natif, et le lexique natif (courbes en jaune-vert / bleu-nuit) ou non-natif (courbes en bleu-ciel / marron). Le paramétrage est dépendant de la longueur des transcriptions.

La Figure 4 montre que l'utilisation des fonctions sigmoïdes dépendantes de la longueur des transcriptions, donne de meilleures résultats que l'utilisation d'une fonction de décision unique (globale), quel que soit le corpus utilisé et quel que soit le lexique associé.

La Figure 5 montre que l'utilisation d'un lexique natif est nécessaire pour le corpus natif. Cependant, pour le corpus non-natif, les deux lexiques donnent des résultats similaires. Les résultats montrent également qu'il est important d'apprendre les fonctions de décision sur les données non-natives pour obtenir des résultats optimaux sur le corpus non-natif. En revanche, l'influence du corpus d'apprentissage semble négligeable pour le traitement de la parole native.

Les meilleurs résultats obtenus pour le corpus natif sont : un taux de *fausse acceptation* de 6.35% avec un taux de *faux rejet* de 9.53%, qui correspondent à une *F* mesure (eq. 1) de 92.031%.

Les meilleurs résultats obtenus pour le corpus non-natif sont : un taux de *fausse acceptation* de 4.88% avec un taux de *faux rejet* de 6.68% (qui correspondent à une F mesure de 94.207%). Par comparaison, lorsque la décision (acceptation/rejet) est effectuée en considérant un seul critère à la fois, nous obtenons pour le critère associé aux phonèmes, F=88.686%, pour celui associé aux trames, F=86.986% et, pour celui associé aux zones de non-parole, F=78.563%.

4 Conclusions

Cet article a étudié le rejet des transcriptions incorrectes de parole non-native dans le cadre de l'apprentissage d'une langue étrangère. Quelques questions se sont alors posées. Comment rejeter les entrées incorrectes tout en tolérant les défauts de prononciations non-natives ? Quels méthodes et critères choisir afin de pouvoir différencier les entrées correctes de celles incorrectes ? L'exploitation de variantes non-natives dans le lexique est-elle bénéfique ? Les fonctions de décision doivent-elles être spécifiques à la longueur des transcriptions ? Sur quel type de corpus (natif ou non-natif) doit-on faire l'apprentissage de paramètres ?

Pour répondre à toutes ces questions, nous avons utilisé deux corpus anglais, l'un prononcé par des locuteurs natifs et l'autre par des non-natifs. De plus, chaque corpus a été découpé en deux, une moitié pour l'apprentissage et l'autre moitié pour les évaluations de performance. Pour l'apprentissage des fonctions de décision dépendantes de la longueur des transcriptions, trois catégories ont été considérées : transcriptions courtes, moyennes et longues. Pour distinguer les transcriptions correctes de celles incorrectes, nous comparons l'alignement contraint par la transcription à vérifier avec l'alignement résultant d'un décodage phonétique non-contraint. Cette comparaison a été faite en exploitant trois critères calculés au niveau de phonèmes, de trames et de zones de non-parole. Ces trois descripteurs sont combinés par une fonction de régression logistique pour fournir la fonction de décision.

Les évaluations effectuées sur ces corpus montrent que :
– l'utilisation de plusieurs fonctions de décision (sigmoïdes) dépendantes de la longueur des transcriptions est plus performante que l'utilisation d'une seule indépendante de la longueur des transcriptions.
– il est important d'apprendre les fonctions de décision sur des données non-natives, en particulier pour le traitement de la parole non-native.
– l'utilisation de variantes de prononciation non-natives dans le lexique des prononciations n'est pas nécessaire pour la tâche de vérification des transcriptions, et même pénalisante si l'on

traite de la parole native.
Le paramétrage optimal amène à des taux de fausse acceptation et faux rejet raisonnables (4.88% et 6.68% pour le corpus de parole non-native).

Remerciements

Les travaux présentés dans cet article font partie du projet ALLEGRO (http ://www.allegro-project.eu/), financé par le programme européen INTERREG IV (http ://www.interreg-fwvl.eu/fr/index.php).

Références

BAZZI, I. et GLASS, J. R. (2000). Modeling out-of-vocabulary words for robust speech recognition. *In ICSLP*, volume 1, pages 401 – 404. 1

BOITE, R. (2000). *Traitement de la parole*. Presses polytechniques et universitaires romandes. 1

DARGNAT, M., BONNEAU, A. et COLOTTE, V. (2010). Intonale : Perception et apprentissage des contours prosodiques en l1 et en l2. `http://mathilde.dargnat.free.fr/INTONALE/intonale-web.html`. 3.1

DREISEITL, S. et OHNO-MACHADO, L. (2002). Logistic regression and artificial neural network classification models. *Journal of Biomedical Informatics*, 35:352 – 359. 2.2

ESKENAZI, M., KE, Y., ALBORNOZ, J. et PROBST, K. (2000). The fluency pronunciation trainer : Update and user issues. *In Proceedings INSTiL2000*. 1

GAROFOLO, J. (2000). An acoustic phonetic continuous speech database. *Speech communication*, 30:95 – 198. 3.1

HENRY, G., BONNEAU, A. et COLOTTE, V. (2007). Tools devoted to the acquisition of the prosody of a foreign language. *In International Congress of Phonetic Sciences - ICPhS 2007*, pages 1593 – 1596. 1

HERRON, D., MENZEL, W., ATWELL, E., BISIANI, R., DANELUZZI, F., MORTON, R. et SCHMIDT, J. A. (1999). Automatic localization and diagnosis of pronunciation errors for second-language learners of english. *In EUROSPEECH*. ISCA. 1

HUNT, A. (1996). `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`. 3.1

JOUVET, D., MESBAHI, L., BONNEAU, A., FOHR, D., ILLINA, I. et LAPRIE, Y. (2011). Impact of pronunciation variant frequency on automatic non-native speech segmentation. *In Language and Technology Conference - LTC'11*, pages 145 – 148. 1

MESBAHI, L., JOUVET, D., BONNEAU, A., FOHR, D., ILLINA, I. et LAPRIE, Y. (2011). Reliability of non-native speech automatic segmentation for prosodic feedback. *In Workshop on Speech and Language Technology in Education - SLaTE 2011*. ISCA. 1, 3.1

WITT, S. et YOUNG, S. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30:95 – 108. 1

YOUNG, S., EVERMANN, G., KERSHAW, D., MOORE, G., ODELL, J., OLLASON, D., POVEY, D., VALTCHEV, V. et WOODLAND, P. (2002). *The HTK Book (for HTK version 3.2)*. Cambridge University Engineering Department. 3.1