Reconnaissance de l'Arabe parlé à partir de modèles acoustiques du Franąais

R. Bayeh (1)(2), S. Lin (2) G. Chollet (2) et C. Mokbel (1)
(1) Université de Balamand
P.O.Box 100 Tripoli, Lebanon
{rania.bayeh, chafic.mokbel}@balamand.edu.lb
(2) Département Traitement du Signal et des Images – ENST
CNRS URA820, 46 rue Barrault, 75634 Paris cedex 13, France
{lin, gerard.chollet}@tsi.enst.fr

Résumé – Abstract

La reconnaissance de la parole multilingue demande une étude de la modélisation acoustique des unités de langue cible en utilisant une ou plusieurs unités des langues sources. Ce papier présente une étude de l'association manuelle ou automatique ''data-driven'' d'unités de cible possibles avec les phonèmes de la langue source. Les unités de cible étudiées sont des mots ou des phonèmes. Des algorithmes pour l'association automatique sont décrits. Tandis que l'association de phonème à phonème est plus pratique, la transcription directe des mots en phonèmes sources fournit de meilleurs résultats. Les expériences montrent que des modèles plus précis de source sont plus appropriés pour déterminer ces associations. Les expériences sont entreprises avec le Français comme langue source et l'arabe comme langue cible.

Multilingual speech recognition pushes to study the acoustic modeling of target language units using one or more source languages' units. This paper presents a study of manual and data driven association of two possible target units with source language's phonemes. The target units studied are words and phonemes. Algorithms for data-driven association are described. While phoneme-to-phoneme association is more practical, word transcriptions provide better results. It has been shown that more precise and rich source models are more suitable to determine those association. Experiments are conducted with French as source language and Arabic as target language.

Keywords – Mots Clés

Reconnaissance de la parole multilingue, Modèles de Markov cachés, Adaptation, Correspondance entre unités acoustiques.
Multilingual Speech Recognition, HMM, MLLR, Acoustical units mapping.

1 Introduction

La recherche en reconnaissance de la parole multilingue a gagné beaucoup d'intérêt ces dernières années. La motivation pour une telle recherche est double, théorique et pratique. Sur le plan théorique, développer des systèmes de reconnaissance de la parole multilingue demanderait un ensemble factorisé et réduit de modèles d'unités acoustiques qui nécessitent de techniques avancées en modélisation acoustique. Sur le plan pratique, nombreuses applications profitent de tels systèmes multilingues. Comme premier pas dans cette direction, ce travail consiste à définir un système automatique de reconnaissance de la parole dans une langue cible basée sur les modèles acoustiques d'une ou plusieurs langues sources. Un avantage majeur d'une telle approche est d'établir des systèmes de reconnaissance de la parole pour des langues ou des dialectes où seulement des bases de données de taille réduite existent. De plus, on s'intéresse au niveau auquel doit se situer la correspondance des unités acoustiques entre les langues source et cible.

Linguistiquement parlant, la représentation phonétique de la langue arabe, qui est notre langue cible, n'est pas unique dû à la contradiction des dialectes utilisés dans les régions différentes malgré l'utilisation de la même langue arabe écrite. Par conséquent, une base de données minimale basée sur un dialecte donné serait insuffisante pour créer et apprendre un modèle acoustique complet. Alors, nous avons conçu un système de reconnaissance pour le dialecte libanais en se basant sur les modèles phonétiques de la langue française. Quant aux unités acoustiques, différents niveaux peuvent être choisis pour chaque langue, «phones», triphones, syllabes ou mots. Dans la présente étude, deux cas sont comparés ; une transcription directe des mots arabes à l'aide des «phones» Français et une correspondance entre les phonèmes.

La correspondance des unités acoustiques des langues source et cible peut être déterminée manuellement ou automatiquement. Pour l'approche automatique, des algorithmes semblables à ceux utilisés pour la détermination des variantes de prononciation dans la modélisation lexicale (Mokbel, 1997) sont employées. Plusieurs occurrences d'apprentissage sont alignées sur une boucle des phonèmes permettant, en se basant sur divers critères, la détermination d'une transcription optimale des mots ou une correspondance entre les phonèmes. La base théorique de ces techniques d'inférence est donnée dans la prochaine section. La section 3 présente les bases de données utilisées dans nos expériences. La section 4 présente les expériences et les résultats obtenus en utilisant des mots en tant qu'unités cibles. La section 5 présente les expériences et les résultats quand des «phones» sont utilisés comme unités cibles. Des tableaux de correspondance de «phones» sont construits manuellement ou automatiquement déterminés à partir des données acoustiques. La pertinence de ces tableaux est étudiée. Finalement, la section 6 fournit des conclusions et des perspectives.

2 Inférence de correspondance entre unités acoustiques

Comme indiqué dans l'introduction, la base théorique de la correspondance automatique entre les unités acoustiques source et cible est semblable à celle utilisée pour la détermination automatique des variantes de prononciation dans la modélisation lexicale (Mokbel, 1997). Supposons que N occurrences $(U_1, \ldots, U_N)$ de la langue cible sont disponibles pour l'apprentissage et que les modèles acoustiques pour la langue source sont $(\lambda_1, \ldots, \lambda_p)$. Nous définissons $(\mu_1, \ldots, \mu_q)$ les q modèles acoustiques pour la langue cible. Supposons qu'un modèle cible peut être exprimé en fonction des modèles acoustiques de source :

$$\mu_i = f_i(\lambda_1, \ldots, \lambda_p) \quad i = 1, \ldots q \qquad \text{(Eq. 1)}$$

Étant donné que, en langue cible, chaque occurrence est exprimée comme succession d'unités acoustiques, le problème général d'optimisation est de trouver $\{\hat{f}_i\}$ comme:

$$\{\hat{f}_i\} = \arg\max_{\{f_i\}} p(U_1,\ldots U_N / \mu_1,\ldots,\mu_q) \qquad \text{(Eq. 2)}$$

Cette optimisation (Eq. 2) dépend de la nature des unités acoustiques et du type de fonctions *{fi}*. Dans le travail actuel les unités de source sont des phonèmes et deux cas sont considérés pour les unités cibles, à savoir les mots et phonèmes.

Si le mot est considéré comme l'unité acoustique cible, la nature de *{fi}* détermine l'algorithme d'optimisation. Si *fi* est la transcription du mot cible *i* en utilisant les phonèmes de source, nous devrions trouver la séquence phonétique qui maximise la vraisemblance sur les séquences d'apprentissage du mot i. Comme en (Mokbel, 1997), les N-meilleures alignements des séquences d'apprentissage du mot *i* sur une boucle de phonèmes de source fournit plusieurs solutions et celle avec la meilleure vraisemblance est choisie.

Quand des phonèmes sont considérés en tant qu'unités cibles, *{fi}* représente l'association de phonèmes source aux phonèmes cible. Comme cette association doit être inférée des occurrences d'apprentissage, les séquences d'apprentissage sont segmentées manuellement en phonèmes. Ces occurrences sont ensuite alignées sur la boucle de phonèmes source. Pour chaque phonème cible les segments acoustiques correspondants sont choisis. Le phonème cible est associé aux *m* phonèmes ayant le chevauchement maximum avec ces segments acoustiques $\{f_i\}$.

Une fois que la correspondance entre unités source et cible déterminée, les paramètres des modèles de source peuvent être adaptés pour mieux décrire la distribution des séquences d'apprentissage. Plusieurs techniques peuvent être employées pour l'adaptation (Odell, 2002). Les expériences entreprises dans ce travail sont limitées à l'adaptation par ''Maximum Likelihood Linear Regression'' (MLLR).

3 Base de données et représentation phonétiques

Deux bases de données sont employées dans nos expériences, la base de données ''Swiss-French Polyphone'' (Chollet, 1996) et une base de données arabe (de dialecte Libanais). La base de données Swiss-French Polyphone se compose approximativement de 10 échantillons de phrase et de 28 mots isolés prononcés par 5000 locuteurs. Deux ensembles de modèles de phone sont appris sur cette base de données. Le premier ensemble, désigné par PL16, se compose de 42 modèles de phone comprenant 2 modèles de silence et 6 modèles de fermeture. Le deuxième ensemble, PL32, est un total de 36 modèles de «phones» avec les modèles de «phones» plosive appris comme un seul phone. Tous les modèles sont appris sur 9938 phrases répétées par 1000 locuteurs (500 hommes, 500 femmes). Pour les deux ensembles, les modèles sont des chaînes de Markov cachés (HMM) gauche-droite de trois états. Les distributions de sortie par état sont un mélange à 16 distributions gaussiennes pour PL16 et à 32 distributions gaussiennes pour PL32.

La base de données arabe, d'autre part, similaire aux bases de données SpeechDat, a été collectée à l'Université de Balamand (UOB) en collaboration avec l'ENST. Cette base de données contient 923 échantillons de mot isolés prononcés par approximativement 50 locuteurs masculins et féminins entre les âges 18 et 25.

HTK (Odell, 2002) a été utilisé pour l'apprentissage de ces modèles. Comme vecteurs de paramètres, des vecteurs de 13 composantes MFCC sont extraits sur des fenêtres de 25.6ms chaque 10ms. Les dérivées du premier et second ordre sont associées aux vecteurs statiques menant à un vecteur de paramètres de 39 composantes.

Pour la représentation phonétique, l'IPA (International Phonetic Alphabet), qui est un outil bien connu pour explorer la similitude phonétique à travers les langues, est employé à ce stade avec SAMPA, la représentation de toutes les transcriptions arabes sur le clavier (IPA).

4 Association au niveau des mots en langue cible

Dans ce premier cas, pour chaque mot cible un modèle est créé en concaténant manuellement ou automatiquement des modèles de phonèmes de source. Ces modèles sont ensuite employés avec des noeuds de silence de début et de fin pour la création des réseaux de mots. Une adaptation sur les occurrences d'apprentissage est enfin conduite. Les sections suivantes décrivent la détermination des séquences et les résultats correspondants.

4.1 Transcription Manuelle

La meilleure représentation phonétique pour chaque mot dans le corpus cible a été déterminée manuellement. Bien que le corpus se compose des échantillons de dialecte libanais seulement, les accents de différentes régions géographiques ont mené à plus d'une transcription par mot. Un exemple est montré dans le tableau 1.

Prononciation Arabe	Transcription Française	
	Manuelle	Automatique
عامل(3amil)	an mm ei ll ai mm ei ll	ff an in mm ai ll ss ff an in mm ee ll ss ff an in mm ee ii ll un

Tableau 1 – Exemples de transcriptions manuelle et automatique

4.2 Transcription Automatique

Pour associer automatiquement chacun des mots cible à une séquence de phonèmes, les séquences d'apprentissage sont alignées sur une boucle de phonèmes élémentaire pour produire les N-meilleurs alignements. Les trois résultats les plus vraisemblables sont choisis pour chaque mot. Ces transcriptions (Tableau 1) sont employées pour établir de nouveaux modèles qui ont été adaptés et évalués. Les figures 2 et 3 montrent les résultats de reconnaissance après différentes itérations d'adaptation par MLLR. La transcription automatique produit de meilleure performance avec un modèle plus simple PL16 de source, alors que nous notons la tendance opposée avec un modèle plus précis PL32. Notre interprétation est que plus de données sont nécessaires pour adapter les modèles PL32.

5 Tableaux de correspondance phonétiques

Le mot est la plus grande unité de langue cible qui offrirait la meilleure précision en utilisant de plus petites unités d'une autre langue. Cela indique que c'est nécessaire de régénérer un nouveau dictionnaire lexicologique pour la langue cible. Il serait plus pratique si une association de plus petites unités est trouvée. Par conséquent, l'association de phonèmes a été étudiée. Cette association a été faite manuellement et automatiquement. Ces tableaux, ainsi que la description lexicologique des mots cible en termes de «phones» de langue cible, sont alors employés pour créer le nouveau modèle en copiant le modèle d'unités de source et en l'étiquetant selon le phone cible qu'il représente. Finalement, de même que dans le cas précédent, l'adaptation MLLR et la reconnaissance sont conduites.

5.1 Correspondance Manuelle

Comme pour la méthode transcription manuelle, la correspondance se base sur l'expertise humaine. Une tableau d'association a été manuellement crée reliant chaque phone arabe à un ou plusieurs «phones» français.

5.2 Correspondance Automatique

Pour créer automatiquement un tableau de correspondance, la première étape était la reconnaissance simple des expressions cible dans l'ensemble d'apprentissage en utilisant une boucle des unités de source. Lors de comparaison des transcriptions résultantes aux originales, le rapport de la correspondance entre chaque unité cible et chaque unité source est calculé. Finalement, l'unité française avec le rapport le plus élevé est associée à chaque phone cible et quelques exemples de la correspondance automatique (comme matrice d'association) sont montrées sur la Fig 1.

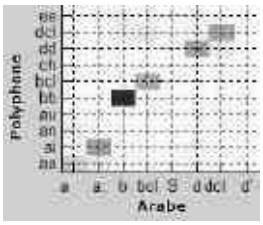
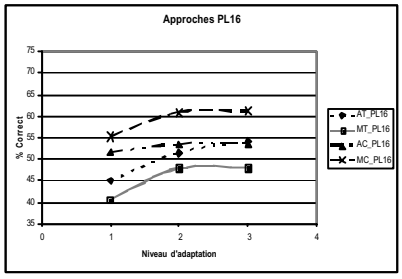
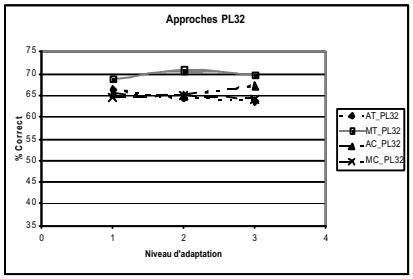


Figure 1 – Correspondance Automatique (phonème à phonème)

Pour vérifier la capacité de généralisation de ces associations nous avons expérimenté l'utilisation des tableaux de correspondance automatique PL16 pour créer des modèles PL32 et vice versa. De tels modèles sont désignés par PL16_32 et PL32_16 respectivement.





Résultats de toutes les approches

Figure 2 – pour PL16 Figure 3 - pour PL32

Les figures 2 et 3 montrent les résultats pour la correspondance de phonèmes et les comparent à ceux obtenus aux transcriptions automatiques de mot. Ici, pour les deux modèles, PL16 et PL32, la correspondance automatique fournit de meilleurs résultats. Notre interprétation est que la méthode de correspondance présente plus de contrainte puisqu'une correspondance de phonème à phonème est choisie. En comparaison, les résultats de l'approche transcription, si manuelle pour PL32 ou automatique pour PL16, sont meilleurs que ceux des approches par correspondance ce qui est prévu intuitivement. Afin d'étudier et comparer les capacités de généralisation de chaque méthode plusieurs expériences sont conduites et les résultats sont rapportés sur les figures 4 et 5. La figure 4 montre les résultats d'identification pour les transcriptions automatiques avec des modèles différents. Il est clair qu'un modèle plus précis, PL32, est préférable pour déterminer les transcriptions qui peuvent être employées avec succès avec un modèle moins précis PL16 plus approprié à l'adaptation avec peu de données.

Concernant la capacité du modèle plus précis pour déterminer de meilleures associations la figure 5 confirme cette tendance. Cependant, ces modèles doivent être maintenus puisque la correspondance de phonèmes est plus contraignante.

6 Conclusions

Dans le cadre général de la reconnaissance de la parole multilingue, ce papier présente un travail sur la correspondance des unités acoustiques entre une langue source et une langue cible. Les phonèmes sont considérés comme les unités acoustiques de la langue source. Deux unités cible sont étudiées ; le mot et les phonèmes. Dans chaque cas, l'association peut être effectuée manuellement ou inférée automatiquement à partir de données acoustiques. Pour la dernière direction les algorithmes sont développés et présentés. La langue source est la langue française tandis que l'arabe (dialecte libanais) est considéré comme langue cible. L'effet de la précision des modèles de langue source est étudié. Les résultats des expériences montrent que l'approche automatique par inférence est généralement plus appropriée. Généralement, il vaut mieux employer des modèles plus précis pour déterminer l'association. Cette association peut être employée avec d'autres modèles moins précis plus appropriés pour l'adaptation.

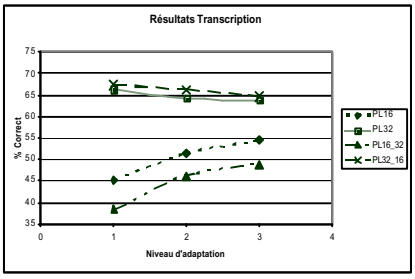


Figure 4 – Résultats des approches de transcription pour tous les modèles

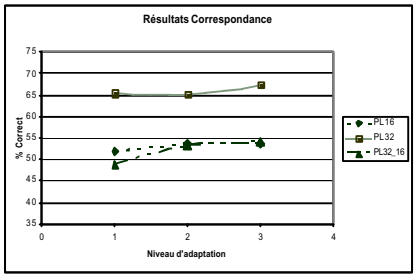


Figure 5 – Résultats des approches de correspondance pour tous les modèles

Comme conclusion finale, nous pouvons dire que bien que la correspondance de phonèmes soit plus appropriée pour établir les modèles de reconnaissance, la transcription des mots fournit de meilleurs résultats. Finalement, plusieurs perspectives existent pour ce travail. Le choix optimal pour les unités acoustiques source et cible doit être déterminé. Établir un ensemble multilingue d'unités acoustiques est une autre perspective.

Remerciements

Ce travail est en partie soutenu par le projet n de CEDRE. (2001 T F 49 /L 42).

Références

E. WONG ET AL (2003), "Multilingual Phone Clustering for Recognition of Spontaneous Indonesian Speech Utilizing Pronunciation Modeling Techniques", *Proc. EuroSpeech'03*, Vol., pp 3133-3136

J. ODELL, D. OLLASON, P. WOODLAND, S. YOUNG, J. JANSEN (2002), "The HTK Book for HTK V3.2", *Cambridge University Press*, Cambridge, UK.

IPA, Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet, Cambridge University Press

H. MOKBEL, D. JOUVET (1997), "Automatic derivation of multiple variants of phonetic transcriptions from acoustic signals," *Proc. EuroSpeech'97*, Vol. 3.

G. CHOLLET ET AL (1996), "Swiss French Polyphone and PolyVar: Telephone Speech Databases to model inter- and intra-speaker variability", *IDIAP-RR 96-0*.