Modélisation acoustique multilingue pour la reconnaissance automatique de la parole non native

Tien-Ping Tan, Laurent Besacier

Laboratoire LIG/GETALP, UMR CNRS 5217
UJF - BP 53, 38041 Grenoble Cedex 9, France
tien-ping@imag.fr, laurent.besacier@imag.fr

ABSTRACT

Automatic speech recognition system performance on non-native speakers is still poor. Non-native speech has a different characteristic compare to native speech as it is influenced by their mother tongue. Getting sufficient non-native speech to create a non-native acoustic model is not always possible. In this paper, we investigate the usage of multilingual acoustic models to adapt the target language acoustic model for non-native speakers. A hybrid of acoustic interpolation and merging is proposed for adapting the target acoustic model by using multilingual acoustic models without requiring the raw corpus. Three types of multilingual resources were tested for adapting non-native speakers: the native language of the speakers, non-native speech from a language different from the target language but spoken by the same native origin, and language close to the target language.

Keywords: non-native ASR, multilingual acoustic modeling

1. INTRODUCTION

Alors que la reconnaissance automatique de la parole atteint désormais des performances souvent satisfaisantes pour les locuteurs natifs, la performance de reconnaissance sur les locuteurs non natifs reste encore insuffisante. Ce problème est un frein au développement des services vocaux dans un monde de plus en plus globalisé.

La parole des locuteurs non natifs a des caractéristiques différentes par rapport aux locuteurs natifs. Pour les apprenants d'une langue, c'est un défi de prononcer les nouveaux phonèmes qui n'existent pas dans leur langue maternelle. Ils ont donc tendance à emprunter les sons de leur langue maternelle. D'autre part, même pour les phonèmes similaires qui existent à la fois dans la langue cible et la langue maternelle du locuteur, les locuteurs non natifs peuvent avoir des difficultés à changer certaines habitudes d'articulation spécifiques à leur langue maternelle.

L'obtention de parole non native pour une langue cible donnée (afin de créer un modèle acoustique qui modélise la parole non native) peut être longue et parfois irréalisable. Par conséquent, les recherches en modélisation acoustique de la parole non native visent plutôt à utiliser la langue maternelle du locuteur ou une quantité limitée de parole non native pour adapter les modèles acoustiques initiaux. Plus précisément, les approches de modélisation acoustique non native peuvent être classées en quatre catégories: la reconstruction de modèle acoustique [1], l'interpolation de modèles acoustiques [2], la fusion (*merging*) de modèles acoustiques [2] ou l'utilisation de techniques plus générales d'adaptation au locuteur [3] [4].

Dans cet article, nous explorons la possibilité d'utiliser différents types de ressources multilingues pour la modélisation acoustique de la parole non native. Trois types de ressources seront notamment étudiés pour l'adaptation de modèles acoustiques à la parole non native : des données correspondant à la langue maternelle du locuteur (L1), des données de parole non native, mais dans une langue différente de la langue cible (L2), et des données correspondant à une langue proche de la langue maternelle du locuteur (L3). Du point de vue des méthodes, nous proposons et testons ici une approche hybride entre interpolation et fusion de modèles acoustiques.

Ce document est organisé de la façon suivante. La section 2 présente notre approche multilingue pour la modélisation acoustique non native. Dans la section 3 nous présentons nos tests expérimentaux et les résultats obtenus. Enfin, la section 4 conclut cet article.

2. MODELISATION ACOUSTIQUE MULTLILINGUE POUR LES LOCUTEURS NON NATIFS

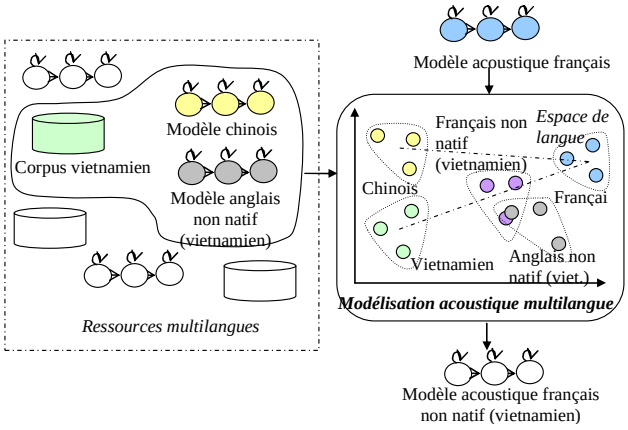

Figure 1. Création d'un modèle non natif français pour des locuteurs vietnamiens, en utilisant des ressources multilingues.

Le concept de modélisation acoustique multilingue pour les locuteurs non natifs consiste à utiliser des ressources multilingues pour adapter la modèle acoustique initial en langue cible. Si nous faisons l'hypothèse qu'aucune donnée de parole non native en langue cible (la langue concernée par le système de reconnaissance automatique de la parole) n'est disponible, alors nous pouvons identifier trois types de ressources multilingues qui peuvent être utiles :

1. des données correspondant à la langue maternelle du locuteur (L1)
2. des données de parole non native, mais dans une langue différente de la langue cible (L2)
3. des données correspondant à une langue proche de la langue maternelle du locuteur (L3)

Par exemple (voir figure 1), si on s'intéresse à l'adaptation d'un système de reconnaissance automatique de la parole du français pour des locuteurs vietnamiens, les ressources multilingues que l'on peut considérer comme potentiellement utiles sont : de signaux de parole ou des modèles acoustiques en vietnamien (L1), des signaux de parole en anglais prononcés par des locuteurs vietnamiens (L2), des signaux de parole ou des modèles acoustiques en chinois mandarin[1] (L3).

2.1. Transfert de phonèmes pour les locuteurs non natifs

Cette étape consiste à déterminer le transfert (ou les confusions) de phonèmes réalisés par les locuteurs non natifs parlant une langue donnée. L'hypothèse ici est que les locuteurs non natifs, lors de l'apprentissage d'une langue (L2), on tendance à utiliser les sons de leur langue maternelle (L1). Le but ici est donc de déterminer les correspondances source / cible selon la langue maternelle des locuteurs non natifs.

Les méthodes pour obtenir ces transferts source / cible sont de deux types : elles peuvent être fondées sur la connaissance ou fondées sur les données. Les approches fondées sur la connaissance consistent à trouver les correspondances source / cible à partir d'études linguistiques existantes [5], de tests de perception, d'analyse phonétique de signaux de parole non native, ou tout simplement en utilisant le tableau de l'Alphabet Phonétique International (API). En revanche, les méthodes fondées sure les données consistent à obtenir automatiquement des confusions source / cible à l'aide de critères de distances (euclidienne, Kullback-Leibler, distances de HMMs [6]), ou en utilisant un décodeur acoustico-phonémique.

Dans cette étude, nous utilisons à la fois une matrice de confusion de phonèmes et l'API pour trouver le phonème source correspondant pour chaque phonème cible, sans utiliser aucune parole non native. La matrice de confusion est créée en alignant les hypothèses du décodeur phonémique en langue source appliqué sur des signaux de paroles en langue cible, et les références obtenues par alignement forcé des ces mêmes signaux avec des modèles acoustiques en langue cible. Le phonème source qu'il est le plus probable de substituer au phonème cible est choisi. Pour déterminer les transferts de phonèmes à partir de l'API en revanche, le phonème source correspondant à chaque phonème cible est déterminé en choisissant le phonème source le plus proche du phonème cible de l'API. Par exemple pour les locuteurs non natifs français d'origine vietnamienne, nous prévoyons que les phonèmes français /p/, /f/, /a/ etc. sont remplacés par les mêmes phonèmes vietnamiens. Pour des phonèmes français qui n'existent pas en vietnamien, par exemple /ʃ/ et /ʒ/, ceux-ci seront remplacés par les phonèmes vietnamiens /s/ et /z/ les plus proches selon l'API. Dans les expériences reportées ici, nous utilisons simultanément les deux méthodes (API et confusion) et choisissons ensuite manuellement le transfert qui nous semble le plus acceptable.

2.2. Modélisation acoustique multilingue pour la parole non native

Une fois les confusions source / cible estimées, l'adaptation du modèle acoustique initial en langue source peut être effectuée. Nous proposons ici une approche hybride d'interpolation et de fusion qui ne s'applique que sur les modèles acoustiques.

L'interpolation de deux modèles source / cible, consiste à trouver, pour chaque gaussienne de chaque état du modèle source, la gaussienne correspondante (dans l'état correspondant) du modèle cible, et à obtenir une distribution gaussienne qui résulte de l'interpolation des deux gaussiennes source / cible. Le nombre de distributions gaussiennes du modèle interpolé reste identique à celui des modèles initiaux source et cible.

En revanche, la fusion de modèles (*merging*), consiste à construire un modèle qui résulte de l'union (pour chaque état) de toutes les gaussiennes source et cible. Le nombre de distributions gaussiennes du modèle issu de la fusion est alors égal à la somme du nombre de gaussiennes des modèles source et cible initiaux.

Nous proposons ici une approche différente qui consiste à proposer un hybride d'interpolation et de fusion. L'approche consiste à associer, pour chaque gaussienne source, une et une seule gaussienne cible. Il est possible, dans cette approche, de traiter les cas où le nombre total de distributions gaussiennes est différent dans le modèle source et dans le modèle cible. Dans ce cas, une gaussienne cible se verra « attribuer » 0 ou plusieurs gaussiennes sources correspondantes. La méthode consiste donc à trouver la distribution gaussienne cible la plus proche pour toutes les distributions gaussiennes source en utilisant par exemple une distance euclidienne. Si la

[1] Le vietnamien et le chinois n'appartiennent pas à la même famille linguistique, mais ce sont deux langues asiatiques tonales qui présentent certaines similarités.

distance entre cette gaussienne cible et la gaussienne source est inférieure à un seuil (cas (1) ci-dessous), alors les gaussiennes sont interpolées ; si en revanche cette distance est supérieure à un seuil (cas (2)), alors nous conservons les deux gaussiennes dans le modèle final (fusion) et les poids associés à ces deux gaussiennes dans le modèle multigaussien final sont ré-estimés suivant l'équation (2). Les gaussiennes cibles auxquelles aucune gaussienne source n'a été associée (cas (3)) sont également fusionnées au modèle multigaussien final.

Etant donné $p_{src} = \{p_{Src,1}, ..., p_{Scr,j}, p_{Src,n}\}$ où p_{Src} représente l'ensemble des distributions gaussiennes source ; étant donné $p_{Tgt,i}$, une distribution gaussienne cible, $p_{Adp,k}$, le modèle adapté avec le poids α, $d()$ la fonction de distance et ω le poids de mélanges gaussiens, on a :

$$p_{Adp,k} = \alpha.p_{Tgt,i} + (1-\alpha).p_{Src,j},\ p_{Src} \neq \phi,\quad d(p_{Tgt,i}, p_{Src,j}) \leq seuil \qquad (1)$$

$$p_{Adp,k} = p_{Scr,j},\ \omega_{Adp,k} = (1-\alpha).\omega_{Scr,j},\ p_{Scr} \neq \phi,\quad d(p_{Tgt,i}, p_{Src,j}) > seuil \qquad (2)$$

$$p_{Adp,k} = p_{Tgt,i},\ w_{Adp,k} = (\alpha).\omega_{Tgt,i},\ p_{Scr} = \phi \qquad (3)$$

3. EXPERIENCES

Nous expérimentons un système de RAP non native en français. Le modèle acoustique initial en langue cible est appris sur le corpus BREF120 [7]. Le système de reconnaissance est développé avec l'outil Sphinx 3 de CMU. Les tests sont effectués sur notre corpus non natif français (NNF) où les phrases sont prononcées par des locuteurs d'origine vietnamienne et chinoise [8]. Ce corpus NNF est très difficile pour la tâche de RAP (taux d'erreurs d'environ 60% !). Les phrases lues sont relatives au domaine du tourisme. Le modèle de langage trigramme est créé en utilisant le texte de journaux *Le Monde*, et ensuite il est interpolé avec un modèle de langage spécifique au tourisme. Le tableau 1 résume tous les corpus utilisés dans les expériences, et le tableau 2 décrit les types de données d'adaptation employées. Des modèles acoustiques indépendants du contexte avec 16 distributions gaussiennes pour le français, le vietnamien, le chinois et l'anglais non natif[2] sont créés pour toutes les expériences reportées ici.

Pour l'approche hybride proposée, la distance euclidienne a été utilisée comme mesure de distance entre les distributions. Les modèles hybrides obtenus ont en moyenne 26 distributions gaussiennes par état. Le seuil qui permet de décider si on interpole ou fusionne (voir équations 1,2 et 3) a été mis à deux fois la distance moyenne entre la distance cible et source.

[2] Celui-ci est créé par adaptation MLLR d'un modèle anglais initial appris sur TIMIT avec des signaux de parole anglaise non native.

Table 1 : Corpus utilisé pour entraînement, adaptation et test

Type	Corpus	Description	Loc.	Heures
Ent.	BREF120	Français	120	100+
Adapt.	VN [9]	Vietnamien	29	15
	CADCC [10]	Mandarin	20	5
	TIMIT [11]	Anglais: sera utilisée avec GMU	630	4
	GMU [12]	Anglais non natif	17	0.14
Test	NNF	Français non natif	10	1

Table 2 : Description des corpus utilisés pour l'adaptation du modèle acoustique initial français

Locuteurs	Corpus	Type de données (voir section 2.)
Vietnamiens	VN	L1
	CADCC	L3 (mandarin)
	TIMIT+GMU	L2 (anglais par vietnamiens)
Chinois	CADCC	L1
	VN	L3 (vietnamien)
	TIMIT+GMU	L2 (anglais par chinois)

Les résultats de la modélisation multilingue pour les locuteurs non natifs, en utilisant les trois types de données L1 (vietnamien), L2 (anglais non natif) et L3 (chinois mandarin) sont présentés sur la figure 2. Les points correspondant à l'abscisse 1.0/0.0 correspondent à la performance du modèle acoustique en français initial, sans aucune adaptation (modèle appris sur BREF120). Les points correspondant à l'abscisse 0.0/1.0 correspondent à l'utilisation de modèles acoustiques en langue source (de type L1, L2 ou L3) pour décoder la parole française non native. Les résultats de l'adaptation du modèle acoustique cible avec des modèles acoustiques de type L1 et L2 (anglais) sont prometteurs. En moyenne, il y a 12,8% et 6.6% d'amélioration relative du taux d'erreur pour les locuteurs vietnamiens et chinois pour l'adaptation avec le modèle de type L1. Les résultats sont meilleurs que les résultats obtenus par fusion simple de modèles (les mêmes tests pour l'approche fusion seule, non reportés en détail ici, montrent une amélioration relative de 11,86% et 5,86% pour le vietnamien et le chinois respectivement, à comparer avec 12.8%/6.6%) tout en permettant d'obtenir des modèles plus petits (moins de distributions). D'autre part, l'amélioration avec L2 pour les locuteurs vietnamiens n'est pas très élevée, parce que seulement 3 minutes de parole de 7 locuteurs est disponibles. Par contre, 5 minutes de parole non native anglaise prononcée par des locuteurs chinois semble aussi efficace que 5 heures de parole de type L1. Une autre observation intéressante du graphique montre que les modèles de type L3 peuvent aussi être utiles pour l'adaptation. En lui accordant un poids approprié, un modèle acoustique chinois semble permettre d'adapter le modèle acoustique français pour les locuteurs vietnamiens et vice versa! Une réduction de 3% est enregistrée lorsque le poids est égal à 0,8 pour les locuteurs vietnamiens et chinois.

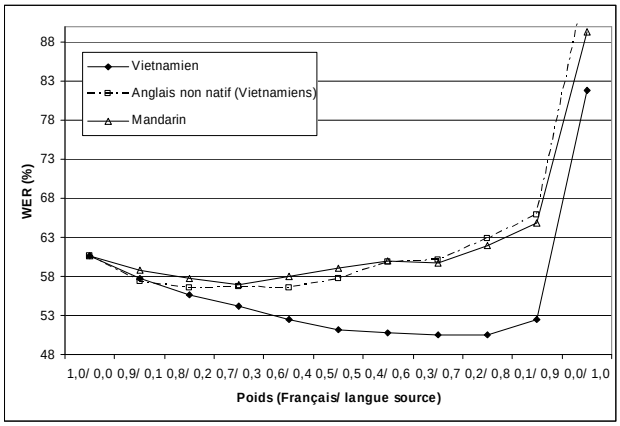

Figure 2 : %WER sur locuteurs non natifs d'origine vietnamienne en utilisant le modèle hybride créé à partir d'un modèle cible en français et de modèles sources de trois types différents (L1 vietnamien, L2 anglais non natif par vietnamiens ou L3 mandarin) pour différents poids.

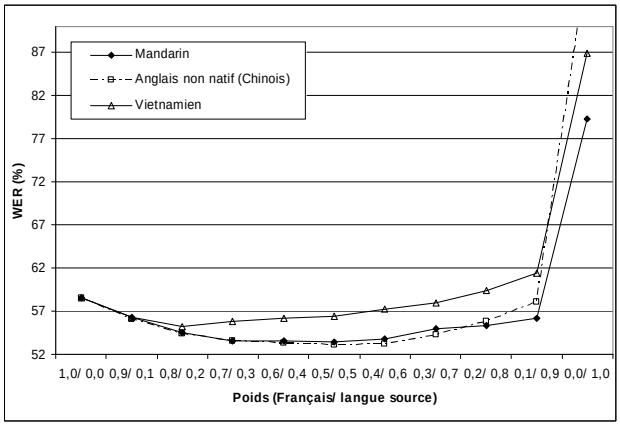

Figure 3 : %WER sur locuteurs non natifs d'origine chinoise en utilisant le modèle hybride créé à partir d'un modèle cible en français et de modèles sources de trois types différents (L1 chinois, L2 anglais non natif par chinois ou L3 vietnamien) pour différents poids.

4. CONCLUSION

Nous avons présenté une approche pour adapter le modèle acoustique cible sans l'utilisation de ressources non natives de la langue cible. Trois types de données peuvent être employées pour l'adaptation non native quand la parole non native cible est indisponible : la langue maternelle du locuteur, toute parole non native du même locuteur natif et les langues proches de la langue maternelle. Parmi ces trois types de parole, la parole non native, même pour différentes langues cible, peut être presque aussi efficace que la langue maternelle du locuteur pour l'adaptation. Avec un poids approprié (bas), une langue proche de la langue maternelle du locuteur peut aussi être utile pour l'adaptation de modèles acoustiques à la parole non native. Plus d'études devront cependant être effectuées afin de déterminer les caractéristiques qui permettent de définir la proximité d'une langue (les langues vietnamiennes et chinoises sont de différentes familles). Une meilleure façon de combiner les différentes langues source peut aussi définitivement améliorer encore le résultat.

BIBLIOGRAPHIE

[1] Y. Liu and P. Fung. Modeling partial pronunciation variations for spontaneous Mandarin speech recognition. *Computer Speech and Language*, volume 17, pages 357-379, 2003.

[2] S.M. Witt. Use of Speech Recognition in Computer-assisted Language Learning. PhD dissertation, Dept. Engineering, University of Cambridge, 1999.

[3] C.J. Leggetter and P. Woodland. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech and Language*, volume 9, pages 171-185, 1995.

[4] Gauvain, Jean -Luc and Lee. Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 2, pages 291-298, 1994.

[5] J. Flege. Second Language Speech Learning Theory, Findings, and Problems. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, pages 233-277, 1995.

[6] B.-H. Juang and L.R. Rabiner. A Probabilistic Distance Measure for Hidden Markov Models. *AT&T Technical Journal*, volume 64(2), pages 391-408, 1985.

[7] L.F. Lamel, J.L. Gauvain and M. Eskénazi. BREF, a Large Vocabulary Spoken Corpus for French. In *Proc. Eurospeech'91*, pages 505-508, 1991.

[8] T.-P. Tan and L. Besacier. A French Non-Native Corpus for Automatic Speech Recognition. In *Proc. LREC'06*, pages 1610-1613, 2006.

[9] V.-B. Le, T. Do-Dat, E. Casteli, L. Besacier and J.F. Serignat. Spoken and written language resources for Vietnamese. In *Proc. LREC'04*, pages 599-602, 2004.

[10] ---, CCC Corpora. http://www.d-ear.com/CCC/corpora.htm, 2005.

[11] W.M. Fisher, G.R. Doddington and K.M. Goudie-Marshall. The DARPA Speech Recognition Research Database: Specifications and Status. In *Proc. of DARPA Workshop on Speech Recognition*, pages 93-99, 1986.

[12] S.H. Weinberger. The Speech Accent Archive. http://accent.gmu.edu/, 2007.