Adaptation rapide de modèles acoustiques compacts

Christophe Lévy, Georges Linarès, Jean-François Bonastre

Université d'Avignon et des Pays de Vaucluse,
Laboratoire Informatique d'Avignon (UPRES 931),
F-84018 Avignon, France

ABSTRACT

In a previous work we presented a new architecture dedicated to embedded speech recognition. It relies on a general GMM, which represents the whole acoustic space, associated with a set of HMM state-dependent probability functions modeled as transformations of this GMM. This work takes advantage of this architecture to propose a fast and efficient way to adapt the acoustic models. The adaptation is performed only on the general GMM model and does not require state-dependent adaptation data. It is also very efficient in terms of computational cost. We evaluate our approach in the voice-command task. This adaptation method achieved a relative error-rate decrease of about 10% even if few adaptation data are available.

Keywords: speech recognition, compact acoustic models, adaptation

1. Introduction

La plupart des systèmes de reconnaissance de parole continue à grand vocabulaire (LVCSR) procèdent à une adaptation non supervisée en utilisant plusieurs passes de décodage. L'idée des différentes méthodes d'adaptation est d'utiliser des transcriptions intermédiaires afin d'adaptater les modèles acoustiques à une personne et/ou un environnement. Habituellement, les fonctions d'adaptation utilisées sont basées sur une MLLR multi-classe[4]. Quelque soit la méthode d'adaptation utilisée, on constate généralement une baisse relative du taux d'erreurs de l'ordre de 10%[1]. Cependant, la précision des fonctions d'adaptation est liée à deux facteurs principaux : le nombre de classes utilisées et la qualité du décodeur (particulièrement durant la première passe de décodage). Le premier facteur (le nombre de classes) dépend majoritairement de la quantité de données d'adaptation disponible, alors que le second (performance du décodeur) nécessite, lui, une grande quantité de calcul et de mémoire.

Ce travail est effectué dans le contexte particulier de la reconnaissance de parole embarquée où les ressources disponibles (tant mémoire que calcul) sont très limitées. De plus, dans ce cadre, l'adaptation des modèles acoustiques est un point essentiel étant donné que le système de reconnaissance peut être utilisé dans de nombreux environnements. Ce dernier point soulève aussi le problème de la quantité de données nécessaires pour adapter les modèles. En effet, le système doit être capable de faire de l'adaptation très rapidement et en utilisant très peu de données propres au nouvel environnement (ou au nouvel utilisateur). Les approches classiques d'adaptation ne semblent donc pas être en mesure d'apporter une réponse adéquate à cette problématique.

Dans un précédent article [5], nous avons proposé une nouvelle architecture, proche des HMM[1] semi-continus, qui utilise un GMM pour représenter l'ensemble de l'espace acoustique (le GMM général) ; chaque état étant ensuite différencié par une simple fonction de transformation appliquée sur ce GMM général. Le GMM général regroupe l'ensemble des informations indépendantes des phonèmes alors que les caractéristiques propres aux phonèmes sont, elles, représentées à l'aide des fonctions de transformation. Cette architecture permet donc de construire des modèles compacts pour la reconnaissance embarquée.

Dans ce papier, nous présentons une méthode d'adaptation dédiée à cette architecture. L'idée principale consiste à adapter uniquement le GMM général sans toucher aux fonctions de transformation. Cela revient à présupposer qu'en déplaçant la référence commune (le GMM général), pour le rapprocher d'un locuteur ou d'un environnement acoustique, on ne modifie pas le pouvoir discriminant des fonctions de transformation. Cette méthode devrait permettre d'adapter l'ensemble des modèles acoustiques avec un très petit nombre de données.

Dans un premier temps, le protocole expérimental général (tâche et corpus) est présenté. La section 3 contient une présentation rapide de l'architecture dédiée à la reconnaissance embarquée. La partie 4 présente en détail la méthode d'adaptation proposée. Enfin, la dernière section présente quelques conclusions et perspectives.

2. Protocole expérimental

Cette section présente le protocole expérimental utilisé pour l'évaluation des méthodes proposées.

2.1. Tâche

Pour estimer les capacités d'adaptation de l'approche proposée, nous nous sommes placés dans le cadre d'une tâche de reconnaissance de commandes vocales. Le taux d'erreur s'exprime en taux d'erreur de commandes (CER).

Dans le cas d'un système de reconnaissance de parole embarqué l'une des principales contraintes reste les ressources (mémoires et calcul) disponibles. Dans ce travail, nous nous sommes principalement intéres-

[1] Hidden Markov Model - modèle de Markov caché

sés à la problématique des ressources mémoires. Deux limites ont été fixées pour la taille des modèles acoustiques :
- *modèle 6k* qui nécessite moins de 6000 paramètres (ce qui correspond au téléphone actuel) et
- *modèle 11k* qui correspond plus à la prochaine génération de téléphone grand public.

2.2. Corpus

Les expériences ont été réalisées avec le corpus VODIS[2] qui est composé de commandes vocales permettant de piloter un système complexe de navigation par satellites à l'intérieur d'un véhicule (système de navigation GPS, téléphone cellulaire et autoradio avec lecteur CD). Ce corpus est composé d'enregistrements effectués par 200 personnes dans deux voitures différentes. Il contient une grande variété de données : lettres, chiffres, commandes vocales, mots épelés, phrases phonétiquement équilibrées...Ces enregistrements sont réalisés avec plusieurs microphones (close-talk et far-talk - micro plus ou moins distants du locuteur). L'environnement acoustique varie suivant les différentes sessions d'enregistrements (les fenêtres sont ouvertes ou non, la radio est allumée ou non, la climatisation est activée ou pas...). Seules les parties contenant les commandes isolées et les phrases phonétiquement équilibrées, enregistrées en condition close-talk (micro près de la bouche), ont été utilisées.

Le corpus a été divisé en trois sous-ensembles :
- VTRAIN : apprentissage. Il contient 2 712 commandes vocales prononcées par 39 personnes. L'apprentissage du modèle général est réalisé en deux temps : un premier GMM est appris en utilisant le corpus BREF[3] (qui contient beaucoup de données mais d'un type très différent de l'application visée) puis, en utilisant les données VTRAIN, il est adapté afin de le rapprocher des données de tests.
- VTEST : évaluation. Il contient 11 136 commandes. Elles sont prononcées par 160 locuteurs qui prononcent les 70 commandes différentes (en moyenne). Aucune adaptation n'est faite avec ces données. Les locuteurs de VTEST sont différents de ceux de VTRAIN (et aussi des locuteurs de BREF).
- VADAPT : adaptation. Il contient 5 phrases phonétiquement équilibrées pour chacun des 160 locuteurs de VTEST. Ces données sont utilisées uniquement pour l'adaptation du GMM général.

3. Modèles acoustiques compacts

Dans [5], nous avons présenté une architecture, pour la représentation des modèles acoustiques, dédiée à la reconnaissance de la parole embarquée. Cette architecture est proche des HMM semi-continus ([10]). L'idée principale est de représenter l'ensemble de l'espace acoustique à l'aide d'un seul GMM (le GMM général) puis de dériver les fonctions de densité de probabilité des états directement depuis le GMM général en appliquant une simple fonction de transformation. Comme illustré par la figure 1, pour un état donné il suffit de sauvegarder les paramètres de la transformation.

Deux familles de fonction de transformation sont présentées (et détaillées) dans [5] : la première est basée uniquement sur une ré-estimation des poids des gaussiennes et la seconde applique une transformation linéraire avant la ré-estimation des poids.

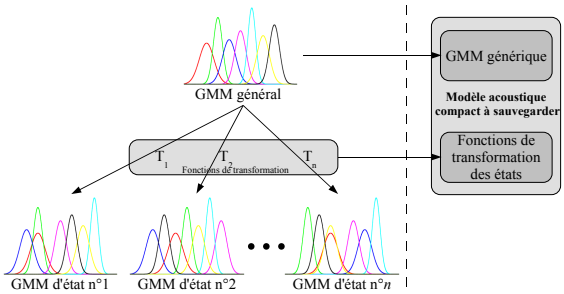


Fig. 1: Architecture pour des modèles acoustiques compacts. Pour un état x, le GMM dépendant de l'état (GMM_x) est obtenu en appliquant la transformation associée à l'état (T_x) au GMM générique.

3.1. Ré-estimation des poids - WRE[2]

Deux critères ont été étudiés pour la ré-estimation des poids.

Le premier, MLE (Maximum Likelihood Estimation), cherche à maximizer la vraisemblance des données en utilisant la règle suivante :

$$\tilde{c}_i = \frac{c_i * Vrais(tr|G_i)}{\sum_{j=1}^{nb_g} c_j * Vrais(tr|G_j)} \quad (1)$$

où c_i correspond au poids *a priori* de la $x^{ième}$ gaussienne, $Vrais(tr|G_i)$ à la vraisemblance de la $x^{ième}$ gaussienne pour la trame tr et nb_g au nombre total de gaussiennes du GMM.

Le deuxième critère utilisé est basé sur la maximisation d'un critère discriminant, MMIE (Maximum Mutual Information Estimation). [8] propose la règle de mise à jour des poids suivante :

$$\tilde{c}_{jm} = c_{jm} * \frac{\gamma_{jm}^{num}}{\gamma_{jm}^{den}} \quad (2)$$

où γ_{jm}^{num} et γ_{jm}^{den} correspondent respectivement aux taux d'occupations estimées des exemples corrects (num) et incorrects (den).

Le tableau 1 présente les résultats obtenus (en terme de taux d'erreurs) en fonction du critère de ré-estimation utilisé.

Aucune diminution du taux d'erreurs n'est constatée et ce quelque soit le critère choisi et la limite pour la taille du modèle acoustique[3]. Cependant, un gain important en terme de calcul est observé avec cette architecture (comparé aux systèmes de reconnaissance classiques).

Tab. 1: Taux d'erreurs obtenu avec l'approche WRE en fonction de la méthode (MLE/MMIE) choisie. 11 136 effectués sur VTEST.

	WRE		HMM
	MLE	MMIE	baseline[5]
modèle 6k	6.05%	5.99%	5.80%
modèle 11k	5.15%	5.15%	4.80%

[2]Weight Re-Estimate

[3]dans [6], des expériences ont montré que dans le cadre de données propres, une baisse signitifactive était constatée.

3.2. *Transformation linéaire - ULT*[4]

Comme le montre le tableau 1, l'approche WRE seule ne permet pas d'obtenir de baisse du taux d'erreurs ; c'est pourquoi nous avons introduit une étape préliminaire qui consiste à appliquer simplement une transformation linéaire sur le GMM général avant l'étape WRE. Cette transformation (appliquée sur la moyenne et la variance) est issue de l'approche LIA_MAP présentée dans [7] et est définie de la manière suivante :

$$\mu_{GMM-état} = \alpha * \mu_{GMM-gnl} + \beta \quad (3)$$

$$\sigma_{GMM-état} = \alpha^2 * \sigma_{GMM-gnl} \quad (4)$$

α est commun pour les deux équations ([7]).

Le tableau 2 montre que l'approche ULT+WRE permet une diminution du taux d'erreurs importante : avec la limite inférieure (modèle de moins de 6k) et l'approche ULT+WRE/MMIE le taux d'erreurs est de 5,11% (soit 12% de moins, en relatif, comparé à la baseline). Dans le cas de la limite supérieures et l'approche ULT+WRE/MLE le taux d'erreurs baisse (toujours en relatif) de plus de 17%.

Tab. 2: Taux d'erreurs obtenu avec l'approche ULT+WRE en fonction de la méthode (MLE/MMIE) choisie. 11 136 effectués sur VTEST.

	ULT+WRE	
	MLE	MMIE
modèle 6k	5.25%	5.11%
modèle 11k	4.01%	4.27%

4. Adaptation

L'approche présentée dans la section précédente modélise les états de manière relative (par la fonction de transformation) à une référence commune (le GMM général). Nous proposons de vérifier si le déplacement de la référence, vers un locuteur par exemple, permet toujours de différencier les états entres eux par une simple transformation. Cette adaptation globale du GMM général s'appuie sur le postulat suivant : si un décalage est constaté entre une unité acoustique indépendante du locuteur et une autre dépendante du locuteur, alors ce même décalage existe probablement entre toutes les unités acoustiques. C'est ce que nous souhaitons mettre en évidence en adaptant le GMM général sans modifier les fonctions de transformation.

4.1. *Description*

Les phrases phonétiquement équilibrées de VADAPT ont été utilisées pour adapter le GMM général afin d'obtenir des GMM généraux propres à chaque locuteurs. Ces phrases sont totalement différentes des commandes à décoder. Cette adaptation MAP a été réalisée suivant la définition de Reynolds ([9]) sur les moyennes des gaussiennes[5]. Le facteur de régulation (ρ) a été fixé à 14 (valeur classique en reconnaissance automatique du locuteur). Dans cette série d'expériences, nous n'avons pas fait d'adaptation en ligne du GMM général ; seule une adaptation du GMM appliquée avant les tests et avec des phrases phonétiquement équilibrées a été réalisée.

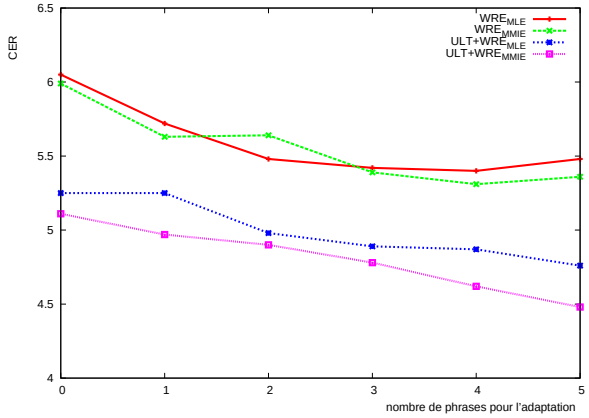
Le processus général se décompose donc en trois étapes :

- phase d'apprentissage : le GMM général et les fonctions de transformation des états sont appris avec VTRAIN ;
- phase d'adaptation : le GMM général est adapté avec le peu de données disponibles dans VADAPT afin d'obtenir un GMM dépendant du locuteur.
- phase de test : les fonctions de transformation apprises durant la phase d'apprentissage sont appliquées directement sur le modèle dépendant du locuteur issue de la phase d'adaptation.

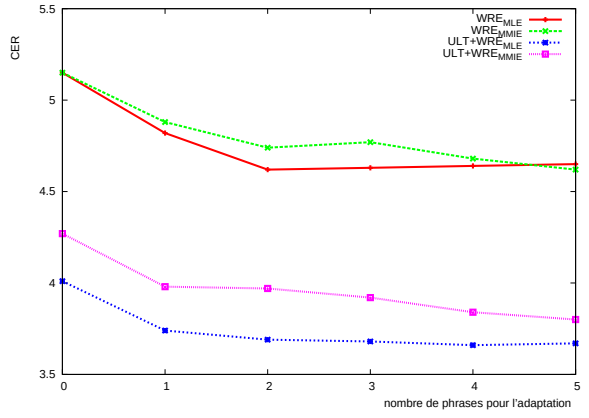
4.2. *Résultats*

Une diminution régulière du CER peut être constatée pour l'ensemble des fonctions de transformation présentées dans cet article. Le graphique 2(a) illustre cela dans le cadre des modèles ayant moins de 11k paramètres. Le graphique 2(b) présente, pour sa part, les différentes valeurs de CER pour le modèle compact. L'évolution du CER est comparable pour les modèles compacts et très compacts.

Nous pouvons constater que, comme attendu, plus le nombre de phrases d'adaptation disponibles est important, plus l'adaptation est performante.



(a) 6k model



(b) 11k model

Fig. 2: Evolution du taux d'erreurs en fonction du nombre de phrases phonétiquement équilibrées utilisées pour l'adapatation du GMM général.

Les tableaux 3 présentent les résultats obtenus pour l'adaptation du GMM général avec l'approche WRE (3(a)) et avec l'approche ULT+WRE (3(b)) en adaptant le GMM général avec les cinq phrases phonétiquement équilibrées dont nous disposons[6]. Une amé-

[4] Unique Linear Transformation

[5] En reconnaissance du locuteur, les systèmes "état de l'art" ne réalisent qu'une adaptation de la moyenne.

[6] il faut noter que pour l'ensemble des résultats présentés l'interval de confiance est de 0,4%.

lioration significative du CER peut être notée et ce quelque soit l'approche utilisée.

Tab. 3: Comparaison des taux d'erreurs pour les approches WRE(3(a)) et ULT+WRE (3(b)) avec et sans adaptation au locuteur (adaptation faite en utilisant 5 phrases phonétiquements équilibrées). 11 136 tests effectués sur le corpus VODIS (l'interval de confiance est de 0,4%).

(a) WRE approach

	without adaptation		with adaptation	
	MLE	MMIE	MLE	MMIE
6k model	6.05%	5.99%	5.48%	5.36%
11k model	5.15%	5.15%	4.67%	4.63%

(b) ULT+WRE approach

	without adaptation		with adaptation	
	MLE	MMIE	MLE	MMIE
6k model	5.25%	5.11%	4.76%	4.48%
11k model	4.01%	4.27%	3.64%	3.80%

En effet, l'approche WRE seule (tableau 3(a)) permet un gain relatif jusqu'à 10%. Le taux d'erreurs du modèle très compact, avec ré-estimation des poids par MMIE, passe de 5,99% à 5,36%, ce qui représente un gain relatif de 10,52%. Le modèle compact obtient des gains similaires (gain relatif de 10,1% pour le modèle compact avec ré-estimation des poids avec MMIE).

Au regard du tableau 3(b), le gain relatif est légèrement meilleur lors de l'utilisation de l'approche ULT+WRE. La diminution du taux d'erreurs se situe entre 9% et 12% (en relatif). Le modèle compact avec ré-estimation MLE obtient le gain le moins élevé : 9,22%, le taux d'erreurs passant de 4,01% à 3,64%. Le modèle très compact avec la même ré-estimation des poids (MMIE) obtient, pour sa part, une réduction relative du taux d'erreurs de 12,33%.

5. Conclusion et perspectives

Cette article traite de la reconnaissance de la parole embarquée. L'utilisation des appareils mobiles (PDA, Téléphone, GPS,...) implique de nouvelles problématiques : puissance de calcul et capacités mémoires notamment. De plus, la reconnaissance de la parole est très liée à l'environnement acoustique : s'il est très différent entre la phase d'apprentissage et la phase de tests, les performances s'en trouvent directement impactées. Ce papier apporte des réponses à ces différentes problématiques. L'architecture présentée permet de réduire de manière importante le nombre de paramètres à sauvegarder. Les performances obtenues par cette architecture sont significativement meilleures (comparées à celles obtenues avec une approche classique) dès lors que le processus complet (ULT+WRE) est utilisé. Une baisse relative du taux d'erreurs, entre 10% et 20%, est observée.

Un des objectifs de cette nouvelle architecture était aussi qu'elle soit facilement adaptable. En effet, une personne doit pouvoir prêter son téléphone ou bien l'utiliser dans son bureau, en voiture puis dans la rue sans que les performances soient dégradées. En tirant partie de l'architecture des modèles acoustiques, nous avons montré que le déplacement de la référence, vers un locuteur par exemple, permet toujours de différencier les états entres eux par une simple transformation. Ceci revient à démontrer que les transformations appliquées au GMM, pour obtenir les modèles des différents états, sont invariantes à une modification du dit GMM. Cette d'adaptation permet d'obtenir une nouvelle baisse relative du taux d'erreurs de l'ordre de 10% (équivalent à ce que l'on trouve dans la littérature pour les systèmes classiques).

Dans un travail futur, nous approfondirons la phase d'adaptation : l'amélioration des fonctions d'adaptation (plus précises qu'une simple adaptation MAP sur les moyennes) et/ou l'adaptation "en ligne" du modèle général devraient permettre d'améliorer encore les performances du système de reconnaissance embarquée.

Références

[1] O. Bellot. *Adaptation au locuteur des modèles acoustiques dans le cadre de la reconnaissance automatique de la parole.* PhD thesis, université d'Avignon, LIA, May 2006.

[2] P. Geutner, L. Arevalo, and J. Breuninger. VODIS - voice-operated driver information systems : a usability study on advanced speech technologies for car environments. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP'2000)*, pages 378–382, Beijing, China, October 2000.

[3] L.F. Lamel, J.L. Gauvain, and M. Eskénazi. BREF, a large vocabulary spoken corpus for French. In *Proceedings of the 2nd European Conference on Speech Communication and Technology (Eurospeech'1991)*, pages 505–508, Genoa, Italy, September 1991.

[4] C.J. Leggetter and P.C. Woodland. Speaker adaptation of continuous density HMMs using multivariate linear regression. In *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP'1994)*, pages 451–454, Yokohama, Japan, September 1994.

[5] C. Lévy, G. Linarès, and J.F. Bonastre. GMM-based acoustic modeling for embedded speech recognition. In *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP'2006)*, Pittsburgh, Pennsylvania, USA, September 2006.

[6] C. Lévy, G. Linarès, P. Nocera, and J.F. Bonastre. *Embedded mobile phone digit-recognition*, chapter 7 in Advances for In-Vehicle and Mobile Systems. H. Abut, J.H.L. Hansen and K. Takeda (Eds.), Springer Science, 2007.

[7] D. Matrouf, O. Bellot, P. Nocera, Linarès, and J.F. Bonastre. Structural linear model-space transformations for speaker adaptation. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech'2003)*, pages 1625–1628, Geneva, Switzerland, September 2003.

[8] D. Povey and P.C. Woodland. Frame discrimination training of HMMs for large vocabulary speech recognition. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1999)*, pages 333–336, Phoenix, Arizona, USA, March 1999.

[9] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10 :19–41, 2000.

[10] S.J. Young. The general use of tying in phoneme-based HMM speech recognisers. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1992)*, pages 569–572, San Francisco, California, USA, March 1992.