Transformations *a priori* et *a posteriori* pour l'adaptation au locuteur

Olivier Bellot, Driss Matrouf, Pascal Nocera

Laboratoire d'Informatique d'Avignon
LIA, Avignon, France
{olivier.bellot, driss.matrouf, pascal.nocera} @lia.univ-avignon.fr

Résumé

The speaker-dependent HMM-based recognizers gives lower word error rates in comparison with the corresponding speaker-independent recognizers. The aim of speaker adaptation techniques is to enhance the speaker-independent acoustic models to bring their recognition accuracy as close as possible to the one obtained with speaker-dependent models. In this paper, we propose a method using test and training data for acoustic model adaptation. This method operates in two steps. The first one performs an *a priori* adaptation using the transcribed training data of the closest training speakers to the test speaker. This adaptation is done with MAP procedure allowing reduced variances in the acoustic models. The second one performs an *a posteriori* adaptation using the MLLR procedure on the test data, allowing mapping of Gaussians means to match the test speaker's acoustic space. This adaptation strategy was evaluated in a large vocabulary speech recognition task. Our method leads to a relative gain of 15% with respect to the baseline system and 10% with respect to the conventional MLLR adaptation.

Les systèmes de reconnaissance de la parole utilisant des modèles acoustiques dépendant du locuteur sont plus performants que ceux basés sur des modèles indépendant du locuteur. Le but des techniques d'adaptation est d'améliorer ces derniers modèles pour s'approcher des performances obtenues avec un modèle dépendant du locuteur. Dans cet article, nous proposons une méthode utilisant les données de test et d'apprentissage pour adapter les modèles indépendant du locuteur. Cette méthode comporte deux types d'adaptation. La première est une adaptation supervisée au moyen de la technique MAP (*Maximum A Posteriori*) et des données d'apprentissage correspondant aux locuteurs les plus proches acoustiquement du locuteur de test. La seconde adaptation est non-supervisée, et est réalisée au moyen de la méthode MLLR (*Maximum Likelihood Linear Regression*) à partir des données de test. Cette stratégie d'adaptation a été évaluée sur le corpus de test de l'AUPELF, ARC B1. Dans ce cadre, notre méthode permet un gain relatif de 15% par rapport au système initial et un gain relatif de 10% par rapport à l'adaptation MLLR habituelle.

1. Introduction

Dans le cadre des systèmes de reconnaissance de la parole indépendant du locuteur, la modélisation des variabilités inter-locuteur nécessite une large population de locuteur. Cette stratégie d'apprentissage entraîne une variance relativement importante dans les modèles acoustiques, réduisant ainsi les capacités du système à differencier les différents phonèmes, et ce particulièrement pour les tâches avec une grande perplexité.

Pour contourner ce problème, deux catégories d'approches ont été proposée. La première consiste en une normalisation de l'espace des vecteurs acoustiques. Dans cette catégorie, nous pouvons y trouver la soustraction du cepstre moyen [2], la normalisation de la longueur du conduit vocal [3], ou enfin une normalisation de l'espace utilisant le critère du maximum de vraisemblance et les MMC (Modèles de Markov Cachés) [4], [5].

La seconde catégorie modifie l'espace des modèles acoustiques. Dans cette catégorie, les méthodes consistent à adapter le modèle indépendant du locuteur à un locuteur spécifique afin d'obtenir un taux de reconnaissance aussi proche que possible de celui obtenu avec un modèle dépendant du locuteur. Dans ce but, les méthodes suivantes ont été proposées : dans [7], la technique d'estimation basée sur le Maximimum *A Posteriori* (MAP) est introduite. Cette méthode vise à obtenir une estimation Bayesienne pour les paramètres des modèles acoustiques en utilisant les données disponibles pour le locuteur de test. Dans [8], le système indépendant du locuteur est adapté au locuteur de test en appliquant une transformation linéaire sur les moyennes des gaussiennes. L'estimation de la transformation est fondée sur le critére du maximum de vraisemblance. Enfin, le modèle compact, technique introduite dans [6], consiste à modéliser séparemment les variations dues au locuteur et à soustraire cette variation des données d'apprentissage. Ceci entraîne une variance réduite des modèles, diminue les recouvrements entre les modèles acoustiques, et permet une meilleure adaptation du modèle au moment de la phase reconnaissance, au moyen d'une transformation de type MLLR.

D'autres schémas d'adaptation sont basés sur le fait que les locuteurs d'apprentissages sont plus ou moins proches acoustiquement du locuteur de test. Par exemple, la technique introduite dans [9] utilise les données d'adaptation pour trouver un sous-ensemble des locuteurs d'apprentissage qui sont les plus proches acoustiquement du locuteur de test. Alors, une transformation linéaire est estimée puis appliquée pour rapprocher l'espace acoustique de chaque locuteur d'apprentissage selectionné à l'espace acoustique du locuteur de test ; enfin, les données ainsi transformées sont utilisées pour adapter le modèle général. La transformation linéaire est estimée en utilisant la procédure MLLR [10].

Dans cet article, nous proposons une méthode [1] qui utilise les données d'apprentissage et les données de test,

en deux étapes : la première consiste en une adaptation utilisant les données d'apprentissage transcrites (adaptation *a-priori* supervisée) ; cette adaptation est effectuée grâce à la procédure MAP. La seconde adaptation est de type MLLR, utilisant les données de test (adaptation *a-posteriori* non-supervisée). Ces deux adaptations ont des buts diffèrents : la première permet essentiellement de réduire la variance des modèles acoustiques tandis que la seconde permet de modifier les moyennes des modèles acoustiques afin d'être le plus proche possible de l'espace acoustique du locuteur de test.

Dans la section 2, nous détaillons la méthode d'adaptation proposée ; nous y décrirons les buts et les stratégies pour l'adaptation *a priori* et *a posteriori* au locuteur. Dans la section 3, nous décrirons deux stratégies de selection des locuteurs d'apprentissage. Dans la section 4, les résultats de plusieurs expériences réalisées sur le corpus ARC B1 seront exposés et commentés. Enfin, dans la section 5 nous verrons les conclusions apportées et les perspectives offertes.

2. Processus d'adaptation

Du fait de la variabilité inter-locuteur, les modèles indépendant du locuteur ont une plus grande variance que les modèles correspondant dépendant du locuteur. En utilisant l'adaptation MLLR, nous adaptons uniquement les moyennes des gaussiennes des modèles. Ainsi, les modèles acoustiques correspondant ont toujours une variance relativement importante, entrainant ainsi un fort recouvrement entre les différentes unités phonétiques. Afin de réduire la variance des modèles, une solution peut être d'utiliser la technique d'adaptation MAP (Maximum *A Posteriori*) [7]. Mais, cette procédure nécessite une quantité de données relativement importante afin de réestimer toutes les variances des gaussiennes. Habituellement, ce sont les données de test qui servent de données d'adaptation, mais celles-ci sont généralement en quantité insuffisante. Notre technique permet de contourner ce problème en utilisant également les données d'apprentissage.

Compte-tenu des contraintes précedemment citées, nous proposons une stratégie générant des modèles acoustiques adaptés avec des variances réduites. Cette adaptation procède en deux étapes (voir Figure 1). La première étape d'adaptation, que nous appelerons *adaptation a-priori*, est réalisée en sélectionnant un sous-ensemble de locuteurs d'apprentissage les plus proches acoustiquement du locuteur de test. Puis, les modèles acoustiques indépendant du locuteur sont adaptés en utilisant les données d'apprentissage transcrites correspondant aux locuteurs ainsi sélectionnés. La première étape d'adaptation est réalisée avec la procédure MAP, qui modifie les moyennes et les variances des gaussiennes comme suit : soit g une gaussienne (appartenant au modèle acoustique indépendant du locuteur) de moyenne μ_g et de variance Σ_g. La nouvelle moyenne $\tilde{\mu}_g$ et la nouvelle variance $\tilde{\Sigma}_g$ de la gaussienne g sont données par :

$$\tilde{\mu}_g = \frac{\eta_g + \tau_g \mu_g}{c_g + \tau_g} \quad (1)$$

$$\tilde{\Sigma}_g = \frac{1}{c_g + \tau_g}[\gamma_g + \tau_g[\Sigma_g + \mu_g \mu_g^{tr}]] - \tilde{\mu}_g \tilde{\mu}_g^{tr} \quad (2)$$

où :

$$c_g = \sum_t c_g(t) \quad (3)$$

$$\eta_g = \sum_t c_g(t) x_t \quad (4)$$

$$\gamma_g = \sum_t c_g(t) x_t x_t^{tr} \quad (5)$$

Le paramètre τ_g est habituellement considéré constant pour toutes les gaussiennes. $c_g(t)$ est la probabilité *a posteriori* de la gaussienne g au temps t, étant données les observations acoustiques $x_{t=1\ldots T}$.

La première étape d'adaptation permet de modifier tous les paramètres des gaussiennes, mais seulement l'adaptation des poids et des variances permet effectivement une amélioration des capacités du modèles pour un locuteur spécifique. Pour réellement adapter les moyennes des gaussiennes pour un locuteur donné, il faut disposer de données de locuteurs suffisemment proches. En utilisant une population relativement peu importante de locuteurs (120 locuteurs), il y a peu de chance de trouver des locuteurs qui correspondent de manière satisfaisante au locuteur de test. Les variations spectrales dues à la variabilité inter-locuteur pour chaque unité phonétique est ainsi réduite par la première adaptation, mais les moyennes des gaussiennes restent inadaptées au locuteur de test ; il est donc nécessaire de procéder à une seconde phase d'adaptation.

La seconde phase consiste en l'adaptation des moyennes des gaussiennes des modèles acoustiques résultant de la première adaptation. Cette seconde adaptation est effectuée au moyen de la procédure MLLR [10] : les données de test sont transcrites en utilisant les modèles acoustiques adaptés par la première étape (les modèles adaptés *a-priori*). Puis, l'alignement trame/état est utilisé pour estimer une transformation linéaire globale qui est appliquée aux moyennes des gaussiennes des modèles acoustiques adaptés précedemment. Nous appelerons cette seconde étape "adaptation *a-posteriori*".

3. Choix du sous-ensemble des locuteurs d'apprentissage

Pour réaliser l'adaptation *a-priori*, nous devons trouver le sous-ensemble de locuteurs d'apprentissage qui soient les plus proches acoustiquement du locuteur de test. Ceci est effectué par le système de reconnaissance automatique du locuteur du LIA, AMIRAL [11], système basé sur des modèles de mixture de gaussiennes (Gaussian Mixture Models, GMM). Pour nos expériences, nous avons utilisé des GMM comportant 128 gaussiennes pour chaque modèle de locuteur. Le processus est le suivant : le système compare tout les locuteurs d'apprentissage au locuteur de test, et classe ces locuteurs du plus près au plus éloignés. Puis, les données transcrites des locuteurs d'apprentissage les plus proches du locuteur de test sont utilisées pour adapter le modèle acoustique indépendant du locuteur au moyen de la procédure MAP.

Une autre stratégie pour choisir les locuteurs d'apprentissage, basée sur des HMM (Hidden Markov Models) plutôt que sur des GMM, a été testée. Tout d'abord,

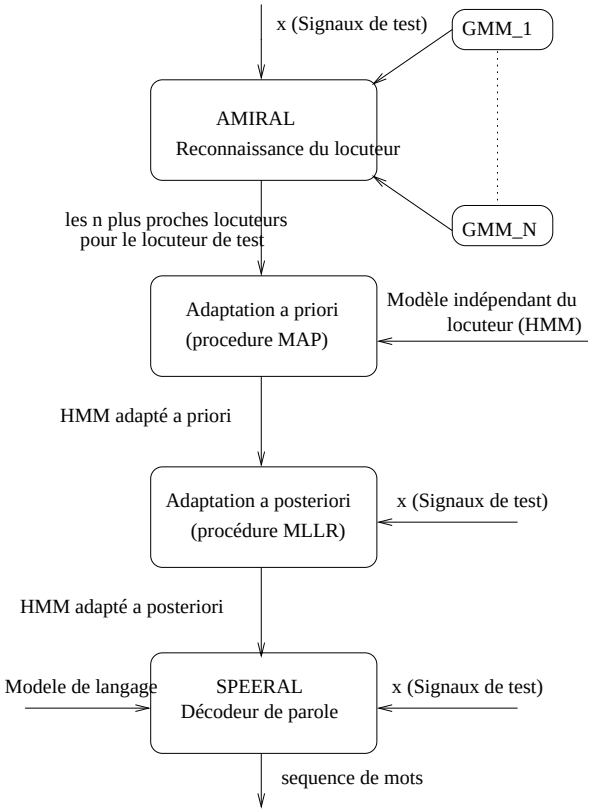

FIG. 1 – *Diagramme du processus d'adaptation.*

nous avons estimé 120 modèles acoustiques dépendant du locuteur (un pour chaque locuteur d'apprentissage). Malheureusement, les données disponibles pour chaque locuteur d'apprentissage ne sont habituellement pas suffisantes pour obtenir une estimation robuste du modèle dépendant du locuteur. Ainsi, nous avons utilisé la méthode MAP [7] pour adapter le modèle indépendant du locuteur à chaque locuteur d'apprentissage, et obtenir 120 HMMs représentant chaque locuteur d'apprentissage. Pour choisir les modèles acoustiques les plus proches des signaux de test, les données de test sont transcrites en utilisant le modèle acoustique indépendant du locuteur. Nous obtenons ainsi un alignement trame/état permettant de calculer la vraisemblance acoustique de chaque modèle dépendant d'un locuteur d'apprentissage. Les n locuteurs ayant les meilleures vraisemblances sont choisis comme étant les locuteurs les plus proches du locuteur de test. Ces deux stratégies ont toujours donné les mêmes cinq plus proches locuteurs.

4. Résultats experimentaux

Dans cette partie, nous présentons les résultats de diverses expériences de reconnaissance. Ces expériences ont été réalisées en utilisant SPEERAL, le système de reconnaissance grand vocabulaire developpé au LIA. Le lexique utilisé contient 20000 mots, et présente un taux de mots hors-vocabulaire de 3.6% pour la tâche choisie. Le modèle de langage utilisé est un trigramme. Le système de base est indépendant du locuteur et du genre. Le modèle acoustique contient 38 phonèmes. Chaque phonème est représenté par un CDHMM (Continous Density Hiden Markov Model, ou Modele de Markov Caché à Densités Continues) de 3 états gauche-droite, indépendant du contexte. Chaque état est une mixture de 64 gaussiennes. Le signal de parole est paramétré en 13 coefficients (12 coefficients mel-cepstraux plus l'energie). Nous utilisons également les dérivées premieres et secondes de ces coefficients ; cela donne un total de 39 coefficients par trame.

Pour estimer les modèles acoustique et linguistique, nous avons utilisé les données d'apprentissage provenant de Bref [3], qui comporte 120 locuteurs (66 femmes et 54 hommes). Les données d'apprentissage contiennent environ 66500 phrases. Les données de test proviennent du corpus ARC B1 de l'AUPELF [2], avec 20 locuteurs et 299 phrases. Les phrases sont des articles publiés dans le journal francais "Le Monde".

L'adaptation *a priori* est réalisée en utilisant les cinq locuteurs d'apprentissage les plus proches acoustiquement du locuteur de test. Cette adaptation est réalisée en utilisant la procédure MAP avec les données d'apprentissage. Dans le tableau 1, nous appelerons cette adaptation *Adapt. 1*. L'adaptation *a posteriori* est réalisée en utilisant le signal de test avec la procédure MLLR. Dans le tableau 1, nous appelerons cette adaptation *Adapt 2*. Les deux adaptations seront comparées avec la procédure MLLR, avec une transformation estimée avec les données de test. Nous appelerons cette adaptation *Adapt. 3*. [1]

	Taux d'Erreur (%)			
	Initial	Adapt. 1	Adapt. 2	Adapt. 3
WER	26.2	25.4	22.4	24.9

Tableau 1 : *Taux d'erreur sur les mots (%) Comparaisons : Adapt. 1 : adaptation a priori avec MAP, Adapt. 2 : adaptation a posteriori avec MLLR sur les modèles obtenus avec Adapt. 1, Adapt. 3 : MLLR sur les modèles acoustiques initiaux.*

Nous pouvons voir que l'adaptation *a priori* utilisant MAP (adapt. 1 dans le tableau 1) n'améliore pas significativement le taux de reconnaissance (seulement 3% de gain relatif par rapport au système initial). Néanmoins, cette étape est importante car, suivie d'une adaptation MLLR, nous observons un gain de 15% par rapport au système initial. Cette amélioration est due au fait que, après l'adaptation MAP, les modèles acoustiques ont une plus petite variance, ce qui permet une adaptation plus efficace des moyennes des gaussiennes.

L'amélioration constatée est très variable d'un locuteur à l'autre. Par exemple, le taux d'erreur d'un locuteur donné était de 41.1% avant adaptation, 40.5% après l'adaptation *a-priori* et enfin 27.9% après l'adaptation *a-posteriori* (1.5% de gain relatif pour l'adaptation *a-priori* contre 32% pour l'adaptation *a-priori* et *a-posteriori*). Pour le test complet, le gain moyen est d'environ 15%. Si, pour certains locuteurs, aucune amélioration n'est constatée, aucun locuteur ne voit son score dégradé. Le constat de l'amélioration plus ou moins importante sui-

[1] La transformation MLLR utilisée dans ces expériences est une transformation globale avec offset

vant les locuteurs peut être expliqué par le fait qu'aucun locuteur d'apprentissage n'est suffisamment proche du locuteur de test.

Dans nos expériences, le gain relatif obtenu par l'adaptation MLLR (une transformation globale) est d'environ 5% par rapport au système initial (de 26.2% d'erreur à 24.9%). Ce gain est trois fois moindre que celui obtenu par la conjonction des adaptations *a-priori* et *a-posteriori*. Le gain relatif obtenu par notre méthode est d'environ 10% par rapport à l'adaptation MLLR conventionnelle.

5. Conclusion

Nous avons présenté une nouvelle méthode d'adaptation des modèles acoustiques utilisant les données de test et d'apprentissage. Cette méthode est composée de deux étapes : la première consiste en une adaptation *a-priori* en utilisant les données d'apprentissage retranscrites et la méthode MAP. La seconde étape, adaptation *a-posteriori*, utilise la procédure MLLR et les données de test. Le but de l'adaptation *a-priori* est de réduire la variance des modèles alors que le but de l'adaptation *a-posteriori* est d'adapter les moyennes des gaussiennes du modèle acoustique afin que celui-ci soit le plus prêt possible de l'espace acoustique du locuteur de test.

Nous avons évalué la méthode proposée dans le cadre de la reconnaissance avec un grand vocabulaire. Les expériences réalisées nous ont emmenés à conclure que la conjonction des adaptations *a-priori* et *a-posteriori* est plus efficace que l'adaptation MAP seule ou que l'adaptation MLLR seule. Le gain relatif obtenu après les deux étapes d'adaptation est d'environ 10% par rapport à la MLLR conventionnelle, de 15% par rapport au système initial, et de 12% par rapport l'adaptation MAP *a-priori* seule.

Dans nos expériences, nous avons choisi arbitrairement les cinq locuteurs les plus proches. Nous avons fait des tests en selectionnant également les dix et quinze locuteurs les plus proches, mais nous n'observons pas de variations significatives des taux de reconnaissance en moyenne. Par contre, nous avons pu observer que le nombre optimal de locuteurs selectionnés diffère suivant les locuteurs de test : certains modèles sont mieux adaptés en choisissant les cinq plus proches locuteurs, d'autres avec dix, d'autres enfin avec quinze. La solution optimale semble donc être de choisir un nombre variable de locuteur suivant le cas, en fonction d'un critère qu'il reste à définir. De plus, nous avons essayé de choisir aléatoirement les locuteurs d'apprentissage servant à l'adaptation *a-priori* ; les résultats sont alors moins bon qu'avant adaptation, validant ainsi notre choix des locuteurs les plus proches.

6. References

[1] D. Matrouf, O. Bellot, P. Nocera, G. Linares, J.-F. Bonastre, "*A Posteriori* and *a Priori* Transformations for Speaker Adaptation in Large Vocabulary Speech Recognition System", Eurospeech 2001, Aalborg.

[2] T. Anastaskos, F. Kubala, J. Makhoul and R. Schwartz, "Adaptation to new microphones using tied-mixture normalization", in *Proc. IEEE Int. Conf. Acoustics, Speech ans Signal Processing*, pp. 433-436, 1994.

[3] H. Eide and H. Gish, "A parametric approach to vocal tract lenght normalization", in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp. 346-349, 1996.

[4] Y. Zhao, "An Acoustic-phonetic-based Speaker Adaptation Technique Improving Speaker-independent Continuous Speech Recongnition", *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 3, pp. 380-394, July 1994.

[5] M. Rahim and B-H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition", in *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 1, pp. 19-30, January 1996.

[6] T. Anastaskos, J. McDonough, R. Schwartz and J. Makhoul, "A Compact Model for Speaker-Adaptive Training", *Proc. ICSLP'96*, pp. 1137-1140, Philadelphia, 1996.

[7] J.-L. Gauvain and C.-H. Lee, "Maximum *A Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", in *IEEE Trans. on Speech and Audio*, MANQUE NUMERO VOLUME ET PAGES, April 1994.

[8] J. R. Bellegarda, P. V. de Souza, A. Nadas, D. Nahamoo, M. A. Picheny and L. R. Bahl, " The Metamorphic Algorithm : A Speaker Mapping Approach to Data Augmentation", *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 3, pp. 413-420, July 1994.

[9] M. Padmanabhan, L. R. Bahl, D. Nahamoo and M. A. Picheny, "Speaker Clustering and Transformation for Speaker Adaptation in Speech Recognition Systems", in *IEEE Transactions on Speech and Signal Processing*, vol. 6, no. 1, pp. 71-77, January 1998.

[10] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", in *Computer Speech and Language*, pp. 171-185, 1995.

[11] C. Fredouille, J.-F. Bonastre and T. Merlin, "AMIRAL : A Block-Segmental Multirecognizer Architecture for Automattic Speaker Recognition", in *Digital Signal Processing*, 2000.

[12] J. Dolmazon, F. Bimbot, G. Adda, J. Caerou, J. Zeiliger, M. Adda-Decker, "Première campagne AUPELF d'évaluation des systèmes de Dictée Vocale", *Ressources et évaluation en ingénierie des langues*, pp. 279-307, 2000

[13] L. F. Lamel et al., "BREF, a Large Vocabulary Spoken Corpus for French", in *EuroSpeech'91*, Genoa, Sept. 1991.