Adaptation acoustique par transformations structurelles estimées dans l'espace des modèles

D. Matrouf, O. Bellot, G. Linares, J-F Bonastre, P. Nocera

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

ABSTRACT

Within the framework of speaker-adaptation, a technique based on tree structure and the maximum a posteriori criterion was proposed (SMAP). In SMAP, the parameters estimation, at each node in the tree is based on the assumption that the mismatch between the training and adaptation data is a Gaussian PDF which parameters are estimated by using the Maximum Likelihood criterion. To avoid poor transformation parameters estimation accuracy due to an insufficiency of adaptation data in a node, we propose a new technique based on the maximum a posteriori approach and PDF Gaussians Merging. The basic idea behind this new technique is to estimate affine transformations which bring the training acoustic models as close as possible to the test acoustic models rather than transformation maximizing the likelihood of the adaptation data. In this manner, even with very small amount of adaptation data, the parameters transformations are accurately estimated for means and variances. This adaptation strategy has shown a significant performance improvement in a large vocabulary speech recognition task, alone and combined with the MLLR adaptation.

1. INTRODUCTION

Un système de reconnaissance automatique de la parole indépendant du locuteur doit être capable de fonctionner quelque soit le locuteur qui utilise ce système. Pour cela, la variabilité inter-locuteur doit être prise en compte lors de l'apprentissage. Ainsi, les systèmes indépendants du locuteur sont typiquement entraînés en utilisant des signaux de parole enregistrés par une population de locuteurs la plus large possible. On obtient alors un modèle acoustique prenant en compte un grand nombre de variabilités acoustiques [8]. Malheureusement, du fait de cette grande variabilité inter-locuteur, les performances d'un système automatique de reconnaissance de la parole indépendant du locuteur sont inférieures aux performances d'un système dépendant du locuteur équivalent. L'idéal serait alors d'utiliser autant que possible des modèles acoustiques dépendant du locuteur. Malheureusement, pour obtenir un système dépendant du locuteur, le besoin de données d'apprentissage pour chaque locuteur de test est tellement important qu'il est impossible, dans la plupart des cas d'utilisation réelle, d'obtenir un modèle pour chaque locuteur de test potentiel.

Les techniques d'adaptation au locuteur transforment les systèmes indépendant du locuteur afin d'obtenir autant que possible un système dont les performances seraient proches de celles obtenues avec un système dépendant du locuteur. Cette adaptation doit utiliser le moins de données d'adaptation possible [8, 6, 1, 7, 2, 11]. La principale difficulté pour l'adaptation au locuteur est d'adapter un nombre important de paramètre avec relativement peu de données d'adaptation. L'adaptation MAP permet une estimation efficace des paramètres du HMM observés dans les données d'adaptation [4], mais ne modifie pas les paramètres qui n'ont pas été observés : l'adaptation MAP est ainsi très locale ; cette technique ne peut donc pas être efficace lorsque les données d'adaptation sont en petite quantité, surtout en adaptation non-supervisée.

Dans le but de contourner ce problème, Shinoda et Lee ont proposé une adaptation structurelle basée sur le critère du maximum *a posteriori* : SMAP (Structural Maximum A Posteriori) [10]. Dans cette technique, les paramètres du modèle acoustique sont supposés suivre une organisation structurelle et cette organisation est représentée à l'aide d'un arbre de classification des états. Les paramètres de transformation pour chaque noeud de l'arbre sont estimés en utilisant l'approche MAP en tenant compte du calcul effectué dans le nœud précédent. La transformation résultante correspondant à chaque paramètre du HMM est une combinaison de toutes les transformations des niveaux supérieurs. Le poids de combinaison entre chaque niveau dépend de la quantité de données d'adaptation présente dans chaque nœud et d'un paramètre fixé *a priori*.

Dans l'adaptation SMAP, l'estimation des paramètres à chaque noeud de l'arbre est basée sur la supposition que le décalage entre le modèle initial et les données d'adaptation peut être modélisé par une gaussienne, appelée "gaussienne de décalage" (*mismatch gaussian*). La moyenne et la variance de cette gaussienne de décalage sont estimées directement sur les données d'adaptation en utilisant le critère du maximum de vraisemblance. De cette manière, la précision de l'estimation de la transformation dépend largement de la quantité de données d'adaptation disponible. Afin d'éviter les mauvaises estimations

de transformations dues à une quantité de données d'adaptation insuffisante, nous proposons une nouvelle méthode d'adaptation basée sur l'approche du maximum *a posteriori* et sur la fusion de gaussiennes. A la différence de techniques telles que SMAP, le but suivi par notre méthode n'est pas de chercher le maximum de vraisemblance des données d'adaptation pour le modèle acoustique, mais plutôt d'estimer des transformations qui rapprochent le modèle acoustique initial du modèle acoustique représentant les données de test, ce modèle acoustique étant estimé en utilisant l'adaptation MAP. De cette manière, même en utilisant peu de données d'adaptation, les paramètres des transformations d'adaptation sont correctement estimés.

Dans cet article, comme dans SMAP [10], nous supposerons que les paramètres des modèles acoustiques sont organisés suivant un arbre de classification contenant toutes les gaussiennes de ce modèle. Chaque noeud de l'arbre représente un sous-ensemble des gaussiennes du modèle acoustique. Toutes les gaussiennes d'un noeud donné partageront la même transformation affine. Cette transformation affine sera sous la forme d'une matrice avec *offset* et aura pour but de réduire le décalage entre les conditions acoustiques d'apprentissage et de test.

Afin d'estimer cette transformation affine, nous proposons une nouvelle technique basée sur la fusion de gaussiennes et sur l'adaptation MAP standard. Cette nouvelle technique est très rapide et permet une bonne adaptation des moyennes et des variances conjointement, et ce même lorsque les données d'adaptation sont en quantités relativement peu importantes et utilisées en mode non-supervisé. A chaque noeud de l'arbre, la transformation est obtenue en combinant trois informations : les données d'adaptation, les paramètres de la transformation du noeud parent et les paramètres adaptés du noeud parent.

Dans la section suivante, nous présentons le processus complet d'adaptation proposé pour ce travail : l'adaptation dans un noeud donné de l'arbre, la combinaison des informations du décalage entre différents niveaux de l'arbre, la fusion de gaussiennes et la construction de l'arbre. La section 3 contient les résultats de plusieurs expériences menées dans le cadre de la reconnaissance automatique de la parole à grand vocabulaire. Enfin, la dernière section (4) contient quelques conclusions, commentaires et perspectives concernant notre nouvelle technique d'adaptation.

2. PROCESSUS D'ADAPTATION

La première partie de cette adaptation est la construction d'un arbre de classification. Chaque nœud de cet arbre représentera un sous-ensemble de gaussiennes du modèle ; la racine de cet arbre contiendra ainsi l'ensemble complet des gaussiennes du modèle. Notons ν un nœud quelconque de l'arbre de classification et $G_\nu = \{g_{m_\nu}, m_\nu = 1 \dots M_\nu\}$ le sous-ensemble de gaussiennes associé au nœud $\nu : g_{m_\nu} = N(\mu_{m_\nu}, \Sigma_{m_\nu})$.

Dans les paragraphes suivants, nous décrivons le processus d'adaptation pour un nœud ν et nous exposerons la stratégie de combinaison sur les différents niveaux de l'arbre.

2.1. Processus d'adaptation dans un nœud

Le but de cette tâche est d'estimer pour chaque nœud ν une transformation affine T_ν (matrice diagonale avec *offset*) partagée par toutes les gaussiennes dans le sous-ensemble G_ν. Cette transformation affine est alors appliquée seulement sur les distributions appartenant à G_ν. Soit $X = \{x_1, x_2, \dots, x_T\}$ décrivant un ensemble donné de T vecteurs acoustiques observés. Notons $\tilde{g}_{m_\nu} = N(\tilde{\mu}_{m_\nu}, \tilde{\Sigma}_{m_\nu})$ la gaussienne obtenue en adaptant la gaussienne $g_{m_\nu} = N(\mu_{m_\nu}, \Sigma_{m_\nu})$ en utilisant l'adaptation MAP standard :

$$\begin{aligned} \tilde{\mu}_{m_\nu} &= \frac{a_{m_\nu} + \tau_{m_\nu}\mu_{m_\nu}}{b_{m_\nu} + \tau_{m_\nu}} \\ \tilde{\Sigma}_{m_\nu} &= \frac{c_{m_\nu} + \tau_{m_\nu}(\Sigma_{m_\nu} + \mu_{m_\nu}\mu_{m_\nu}^{tr})}{b_{m_\nu} + \tau_{m_\nu}} - \tilde{\mu}_{m_\nu}\tilde{\mu}_{m_\nu}^{tr} \end{aligned}$$

où : $a_{m_\nu} = \sum_t \gamma_{m_\nu t} x_t$, $b_{m_\nu} = \sum_t \gamma_{m_\nu t}$, $c_{m_\nu} = \sum_t \gamma_{m_\nu t} x_t x_t^{tr}$, et $\gamma_{m_\nu t}$ est la probabilité *a posteriori* de la gaussienne g_{m_ν} au temps t, calculée sur toutes les observations acoustiques $x_{t=1\dots T}$. Cette probabilté est obtenue en utilisant le moteur de reconnaissance de la parole avec le modèle acoustique initial. Le paramètre τ_{m_ν} est habituellement choisi constant pour toutes les gaussiennes.

Notons $\tilde{G}_\nu$ le sous-ensemble des gaussiennes adaptées à l'aide de la procédure MAP dans le nœud ν : $\tilde{G}_\nu = \{\tilde{g}_{m_\nu}, m_\nu = 1 \dots M_\nu\}$. Notons $g_\nu = N(\mu_\nu, \Sigma_\nu)$ et $\tilde{g}_\nu = N(\tilde{\mu}_\nu, \tilde{\Sigma}_\nu)$ les gaussiennes obtenues en fusionnant respectivement les gaussiennes des sous-ensembles G_ν et $\tilde{G}_\nu$ (voir section 2.2). La transformation affine T_ν est alors estimée pour rapprocher la gaussienne g_ν vers la gaussienne $\tilde{g}_\nu$. Chaque gaussienne $g_{m_\nu} = N(\mu_{m_\nu}, \Sigma_{m_\nu})$ est alors adaptée comme suit :

$$\mu'_{m_\nu} = \tilde{\Sigma}_\nu^{\frac{1}{2}} \Sigma_\nu^{-\frac{1}{2}} (\mu_{m_\nu} - \mu_\nu) + \tilde{\mu}_\nu \quad (1)$$

$$\Sigma'_{m_\nu} = \tilde{\Sigma}_\nu \Sigma_\nu^{-1} \Sigma_{m_\nu} \quad (2)$$

où μ'_{m_ν} et Σ'_{m_ν} sont respectivement les paramètres adaptés de μ_{m_ν} et Σ_{m_ν}.

Cette procédure d'adaptation peut être effectuée itérativement. Nous avons observé expérimentalement que la vraisemblance des données d'adaptation augmente à chaque itération.

2.2. Processus de fusion de gaussiennes

Le processus de fusion de gaussiennes est effectué paire par paire jusqu'à obtenir une seule gaussienne. Dans ce travail, la fusion de deux gaussiennes est basée sur le critère du minimum de perte de vraisemblance.

Notons $G = \{g_1, g_2, \dots, g_n\}$ un sous-ensemble de gaussiennes qu'il faut fusionner de manière à obtenir une gaussienne représentant le sous-ensemble

G. Soient deux gaussiennes $g_i = N(\mu_i, \Sigma_i)$ et $g_j = N(\mu_j, \Sigma_j)$ appartenant à G. Notons c_i et c_j les poids associés respectivement aux gaussiennes g_1 et g_j. La gaussienne $g = N(\mu, \Sigma)$, résultat de la fusion de g_i et de g_j est obtenue en utilisant les formules suivantes :

$$\mu = \frac{c_i\mu_i + c_j\mu_j}{c_i + c_j}$$

$$\Sigma = \frac{c_i\Sigma_i + c_j\Sigma_j + \frac{c_i * c_j}{c_i + cj}(\mu_i - \mu_j)(\mu_i - \mu_j)^{tr}}{c_i + c_j}$$

Le poids c de la nouvelle gaussienne g est la somme des deux poids c_i et c_j. Les deux gaussiennes g_i et g_j du sous-ensemble G sont remplacées par la gaussienne g. Nous répétons cette procédure de fusion jusqu'à avoir une seule gaussienne représentant le sous-ensemble G. Le poids initial c_{m_ν} associé à la gaussienne g_{m_ν} est la somme des probabilités *a posteriori* pour tous les vecteurs acoustiques observés : $c_{m_\nu} = \sum_t \gamma_{m_\nu t}$.

2.3. Adaptation utilisant l'arbre de classification

Dans la section 2.1, nous avons exposé l'estimation de la transformation affine T_ν associée au nœud ν. L'estimation de la transformation T_ν était seulement basée sur les gaussiennes appartenant au nœud courant. Nous allons voir maintenant comment cette transformation est estimée en utilisant l'arbre de classification afin d'adapter l'ensemble du modèle acoustique.

Soit $p(\nu)$ le nœud père du nœud ν. Soit g_ν et $g_{p(\nu)}$ les deux gaussiennes obtenues en fusionnant respectivement les ensembles de gaussiennes G_ν and $G_{p(\nu)}$ (il s'agit des gaussiennes avant adaptation). Notons $\tilde{g}_\nu$ et $\tilde{g}_{p(\nu)}$ les gaussiennes obtenues en fusionnant les gaussiennnes appartenant respectivement aux nœuds $\tilde{G}_\nu$ and $\tilde{G}_{p(\nu)}$ (il s'agit des gaussiennes obtenues avec l'adaptation MAP dans les nœuds ν et $p(\nu)$).

Nous fusionnons les gaussiennes g_ν et $g_{p(\nu)}$ pour obtenir la gaussienne $g_\nu^{p(\nu)} = N(\mu_\nu^{p(\nu)}, \Sigma_\nu^{p(\nu)})$ et nous fusionnons les gaussiennes $\tilde{g}_\nu$ et $\tilde{g}_{p(\nu)}$ pour obtenir la gaussienne $\tilde{g}_\nu^{p(\nu)} = N(\tilde{\mu}_\nu^{p(\nu)}, \tilde{\Sigma}_\nu^{p(\nu)})$. Dans le processus de fusion, le poids associé à la gaussienne du nœud père $p(\nu)$ est un paramètre fixé et le poids associé à la gaussienne du nœud ν est la somme des poids associés aux gaussiennes de ce nœud ($\sum_m c_{m_\nu} = \sum_m \sum_t \gamma_{m_\nu t}$). La transformation affine T_ν est alors estimée comme celle qui rapproche la gaussienne $g_\nu^{p(\nu)}$ vers la gaussienne $\tilde{g}_\nu^{p(\nu)}$. Chaque gaussienne est alors adaptée comme suit :

$$\mu'_{m_\nu} = (\tilde{\Sigma}_\nu^{p(\nu)})^{\frac{1}{2}}(\Sigma_\nu^{p(\nu)})^{-\frac{1}{2}}(\mu_{m_\nu} - \mu_\nu^{p(\nu)}) + \tilde{\mu}_\nu^{p(\nu)}$$

$$\Sigma'_{m_\nu} = (\tilde{\Sigma}_\nu^{p(\nu)})(\Sigma_\nu^{p(\nu)})^{-1}\Sigma_{m_\nu}$$

où μ'_{m_ν} et Σ'_{m_ν} sont les paramètres adaptés respectivement de μ_{m_ν} and Σ_{m_ν}. De cette manière, les paramètres de la transformation résultante est une combinaison des transformations estimées aux nœuds précédents avec la transformation du nœud courant. Le poids servant à la combinaison des transformations entre deux niveaux est ainsi proportionnel à la quantité de données observées pour chaque nœud.

2.4. Construction de l'arbre de classification

Dans le travail présenté, nous avons utilisé des arbres binaires. Nous supposons que toutes les gaussiennes d'un état donné appartiennent à la même classe. Chaque nœud de l'arbre sera ainsi un ensemble d'états constituant un ensemble de gaussiennes. Pour l'étape de classification, chaque état sera représenté par une seule gaussienne obtenue par fusion de toutes les gaussiennes de l'état. Le critère du minimum de perte de vraisemblance sera utilisé pour la classification. L'arbre est construit de *haut-en-bas* : la construction débute à partir de la racine regroupant l'ensemble des états, puis cet ensemble est coupé en deux ; ce processus est appliqué jusqu'aux feuilles [1]. Pour obtenir une classification optimale, il faudrait tester les 2^{n-1} sous-ensembles possibles pour chaque nœud. Pour des raisons évidentes de complexité, nous utilisons à la place une procédure basée sur les *k-plus proches voisins*.

3. RÉSULTATS

Dans cette partie, nous présentons les résultats de diverses expériences de reconnaissance. Ces expériences ont été réalisées en utilisant SPEERAL [9], 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. Le modèle acoustique contient 38 phonèmes. Chaque phonème est représenté par un CDHMM (*Continous Density Hidden Markov Model*, ou Modèle de Markov Caché à Densités Continues) de 3 états gauche-droite, dépendants du contexte. Chaque état est une mixture de 64 gaussiennes. Le signal de parole est paramétré en 13 coefficients (12 coefficients melcepstraux plus l'énergie). Nous utilisons également les dérivées premières 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 [5], 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 [3], avec 20 locuteurs et 299 phrases. Les phrases sont des articles publiés dans le journal français "Le Monde".

Dans nos expériences, nous avons utilisé deux arbres binaires de six niveaux : un pour le modèle acoustique des hommes, l'autre pour le modèle acoustique des femmes. Ces arbres de classification ont été calculés *a priori* (c'est-à-dire avant tout processus de reconnaissance et d'adaptation). Les vecteurs de moyennes et de covariances ont été adaptés, l'adaptation ayant été réalisée locuteur par locuteur dans le mode non-supervisé.

Nous noterons la technique proposée SMAPGM (pour *Structural Adaptation using MAP and Gaussians*

[1] chaque feuille contenant un seul état

Merging techniques ou *Adaptation Structurelle utilisant MAP et la fusion de Gaussiennes*). Dans la table 1, nous pouvons voir que la technique SMAPGM donne un gain relatif d'environ 16% par rapport au système initial. Il est important de noter que les améliorations des adaptations MLLR et SMAPGM peuvent être cumulées. Au final, en effectuant l'adaptation MLLR suivie de l'adaptation SMAPGM, le gain relatif est d'environ 18% par rapport au système initial et en effectuant l'adaptation SMAPGM suivie de l'adaptation MLLR le gain relatif cumulé est d'environ 19,5%. Dans ces expériences, nous avons constaté que les gains les plus importants ont été obtenues sur les tests des locuteurs ayant un taux d'erreur initial important.

Tab. 1: Taux d'erreur sur les mots (%) pour la reconnaissance de la parole avec des modèles dépendant du genre avec diférentes techniques d'adaptation. SMAPGM désigne la méthode proposée : Structural adaptation using MAP and Gaussians Merging technique

Techniques d'adaptation	Taux d'erreur(%)		
	Homme	Femme	Moy.
Base	21,2	21,0	21,1
SMAPGM	18,0	17,7	17,8
SMAPGM+MLLR	16,6	17,4	17,0
MLLR+SMAPGM	17,1	17,5	17,3

Nous avons également réalisé ces mêmes expériences avec un système ayant un meilleur lexique et un meilleur modèle de langage. Le taux d'erreur initial est alors de 19,5%. Après l'adaptation SMAPGM, le taux d'erreur est de 16,6% (soit un gain relatif de 13% par rapport au nouveau système initial, au lieu de 16% pour les expériences menées avec l'ancien système). Lorsque l'adaptation MLLR est suivie de l'adaptation SMAPGM, le taux d'erreur tombe à 15,9% (un gain relatif de 16% par rapport au nouveau système initial, au lieu de 19,5% avec l'ancien système). De cette nouvelle expérience, nous pouvons conclure que le gain relatif obtenu en utilisant l'adaptation SMAPGM semble être plus important lorsque le système initial donne un taux d'erreur plus élevé (ce qui est également observé pour les locuteurs ayant un taux d'erreur élevé).

4. CONCLUSION

Nous avons présenté une nouvelle technique d'adaptation des modèles acoustiques. Cette méthode d'adaptation est basée sur l'adaptation MAP et sur la fusion de gaussiennes. Son efficacité a été confirmée par des expériences en mode non-supervisé sur une tâche de reconnaissance de la parole avec un grand vocabulaire. Le gain relatif constaté par rapport au système initial est de 16%. De plus, il est important de noter que les gains de notre technique peuvent être cumulés avec la technique MLLR ; ainsi, en appliquant notre technique suivi de la MLLR, nous obtenons un gain relatif de 19.5%. Nous avons également montré que notre technique permet de meilleures performances d'adaptation que la technique SMAP.
Afin d'optimiser notre technique, des études portant sur la profondeur d'exploration de l'arbre restent à effectuer [2] (les expériences présentées dans cet article ont été menées en fixant la profondeur de l'arbre). D'autres paramètres importants pour cette technique d'adaptation devraient également être étudiés : les poids servant à la combinaison des informations entre un nœud donné et son parent, ou encore le poids utilisé lors de l'adaptation MAP. Enfin, des études comparatives entre différentes méthodes de construction de l'arbre de classification permettraient de choisir un arbre qui serait optimal pour notre méthode d'adaptation.

Références

[1] T. Anastaskos, J. McDonough, R. Schwartz, and J. Makhoul. A compact model for speaker-adaptive training. In *Proceedings ICSLP*, 1996.

[2] V. V. Digalakis and L. G. Neumeyer. Speaker adaptation using combined tranformation and bayesian methods. *IEEE Trans. on Speech and Audio Processing*, July 1996.

[3] J. Dolmazon, F. Bimbot, G. Adda, J. Caerou, J. Zeiliger, and M. Adda-Decker. Première campagne aupelf d'évaluation des systèmes de dictée vocale. *Ressources et évaluation en ingénierie des langues*, 2000.

[4] J.-L. Gauvain and C.-H. Lee. Maximum *A Posteriori* estimation for multivariate gaussian mixture observations of markov chains. *IEEE Trans. on Speech and Audio*, 1995.

[5] L. F. Lamel, J.L. Gauvain, and M. Eskenazi. Bref, a large vocabulary spoken corpus for french. In *Proceedings Eurospeech*, Genoa, 1991.

[6] C.-H. Lee. On stochastic feature and model compensation approaches to robust speech recognition. *Speech Communication*, 1998.

[7] C.J. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, 1995.

[8] D. Matrouf, O. Bellot, P. Nocera, G. Linares, and J.-F. Bonastre. *A Posteriori* and *a Priori* transformations for speaker adaptation in large vocabulary speech recognition system. In *Proceedings Eurospeech*, Aalborg, Danmark, 2001.

[9] P. Nocera, G. Linares, D. Massonié, and L. Lefort. Phoneme lattice based A* search algorithm for speech recognition. In *Proceedings TSD2002*, Brno, Sept. 2002.

[10] K. Shinoda and C.-H. Lee. Unsupervised adaptation using structural bayes approach. In *Proceedings IEEE ICASSP*, 1998.

[11] O. Siohan, T. A. Myrvoll, and C.-H Lee. Structural maximum a posteriori linear regression for fast hmm adaptation. In *Workshop on automatic speech recognition : challenges for new millenium*, Sept 2000.

[2] un critère pertinent pourrait être par exemple la quantité de donnée d'adaptation : plus les données seraient en quantité importante, plus l'arbre serait parcouru en profondeur