Comparaison de *SMLLR* et de *SMAP* pour une adaptation au locuteur en utilisant des modèles acoustiques markoviens

Fabrice LAURI, Irina ILLINA, Dominique FOHR

LORIA/INRIA
B.P. 239 54506 Vandoeuvre-Lès-Nancy
Mél : {Fabrice.Lauri, Irina.Illina, Dominique.Fohr}@loria.fr

ABSTRACT

In this paper, two adaptation schemes are presented : *SMAP* and *SMLLR*. Both methods update the parameters of the acoustic models of a speaker-independant system in order to improve its performances for a new speaker. We experimented *SMAP* and *SMLLR* to HMMs of the ESPERE engine in the batch mode and in the unsupervised incremental mode. The HMMs were learned on the *Resource Management (RM)* corpus. Results of the batch adaptation show a greatest efficiency of *SMAP*. For the unsupervised incremental adaptation, *SMLLR* is more powerful than *SMAP*, according to the incremental scheme we choose.

1. INTRODUCTION

Tous les systèmes de reconnaissance automatique de la parole (SRAP) actuels utilisent une phase préliminaire d'adaptation, pour améliorer leurs performances lorsqu'ils sont utilisés dans des conditions acoustiques qui diffèrent des conditions acoustiques d'apprentissage. L'emploi des techniques d'adaptation permet de mieux prendre en compte les caractéristiques d'un nouveau locuteur (adaptation au locuteur) ou les caractéristiques acoustiques d'un nouvel environnement (adaptation à l'environnement). Dans la suite de cet article, nous nous plaçons essentiellement dans le cadre de l'adaptation au locuteur.

Les techniques d'adaptation au locuteur peuvent opérer aux différents niveaux d'un SRAP : au niveau du signal de parole, au niveau des modèles acoustiques, au niveau des modèles de langage ou au niveau du dictionnaire de prononciation. Nous nous sommes intéressés ici uniquement aux techniques d'adaptation des modèles acoustiques.

Les techniques d'adaptation des modèles acoustiques permettent d'estimer les paramètres des modèles acoustiques à partir d'un corpus d'adaptation contenant quelques phrases prononcées par le locuteur de test. Ces techniques nécessitent une transcription de chaque phrase d'adaptation pour pouvoir modifier les paramètres des modèles acoustiques. Cette transcription est soit donnée par l'utilisateur (adaptation supervisée), soit obtenue à l'issue d'une phase de reconnaissance par le système (adaptation non supervisée).
Par ailleurs, l'adaptation peut être réalisée soit en mode par lot[1], soit en mode incrémental. En mode par lot, l'adaptation est réalisée à partir du système indépendant du locuteur une fois qu'un certain nombre de phrases d'adaptation sont récoltées. En mode incrémental, l'adaptation s'effectue à partir du système précédemment adapté en utilisant à chaque fois très peu de données d'adaptation.

Plusieurs techniques ont été proposées selon le mode d'adaptation suggéré par le cadre applicatif du SRAP. La plupart de ces techniques sont dérivées de *Maximum A Posteriori (MAP)* [5] ou de *Maximum Likelihood Linear Regression (MLLR)* [9]. La version standard de *MLLR* n'adapte que la moyenne des gaussiennes des modèles acoustiques en utilisant une transformation linéaire estimée au maximum de vraisemblance *(Maximum Likelihood)*. Gales [4] permet une adaptation à la fois des moyennes et des variances. Afify [1] propose de contraindre la structure de la transformation pour assurer une structure probabiliste stable des paramètres dans le cas où les données d'adaptation sont dispersées ou en faible quantité. Digalakis [2] propose une méthode spécifiquement incrémentale et non supervisée [2]. Shinoda [11], [12] et Myrvoll [10] estiment les paramètres des transformations au *Maximum A Posteriori (MAP)*, en introduisant une structure hiérarchique des probabilités a priori sur les paramètres.

Dans cet article, nous nous sommes intéressés à deux méthodes existantes d'adaptation des modèles acoustiques : *Structural Maximum Likelihood Linear Regression (SMLLR [3])* [8] et *Structural Maximum A Posteriori (SMAP)* [11], [12]. Nous avons développé ces méthodes pour le système de reconnaissance ESPERE. Nous les avons évalué sur le corpus *RM* pour une adaptation au locuteur en mode par lot et en mode incrémental non supervisé. La technique proposée par Digalakis [2] a été utilisée pour cumuler les statistiques pour le mode incrémental. Sur la base des expériences réalisées, nous préciserons pour chaque mode d'adaptation quelle technique est la mieux adaptée pour permettre à un SRAP d'améliorer de manière significative ses performances.

La suite de cet article est organisée de la manière suivante. Dans un premier temps, nous exposons les deux méthodes d'adaptation *SMLLR* et *SMAP*. Nous indiquons ensuite les expériences réalisées avec ces deux méthodes, ainsi que les résultats obtenus. Nous concluons enfin sur ces résultats et donnons nos perspectives de recherche envisagées.

[1] Batch adaptation.

[2] Nous donnons au paragraphe 3.1 quelques précisions sur cette technique.

[3] Nous avons pris la liberté d'utiliser l'acronyme *SMLLR* pour indiquer qu'il s'agit de la version structurelle de *MLLR*, même s'il n'a jamais été employé.

2. ADAPTATION DES MODÈLES ACOUSTIQUES

2.1. Adaptation structurelle

Nous supposons par la suite que les modèles acoustiques sont représentés par des modèles markoviens cachés (HMMs). Pour pouvoir donner une estimation fiable des paramètres des transformations quelque soit la quantité de données d'adaptation disponible, *SMLLR* et *SMAP* utilisent une structure d'arbre. Chaque noeud de l'arbre contient plusieurs gaussiennes issues des modèles d'apprentissage et acoustiquement proches au sens d'une certaine mesure de distance. Les gaussiennes des noeuds feuilles sont adaptées en utilisant la même fonction. La profondeur de l'arbre varie avec la quantité de données d'adaptation disponible. Si elle est faible, l'arbre est peu profond, ce qui permet une adaptation globale des paramètres des gaussiennes. Plus elle est grande, plus l'arbre devient profond, permettant donc d'adapter de manière plus précise les paramètres des gaussiennes. Dans le cas extrême où les données d'adaptation sont en très grand nombre, une adaptation locale de chaque gaussienne (représentant chacune un noeud feuille) peut alors être réalisée.

2.2. SMLLR

Nous considérons ici que seules les moyennes des gaussiennes des modèles acoustiques sont adaptées. *SMLLR* [8] suppose que le vecteur moyenne adapté de chaque gaussienne est une combinaison linéaire des vecteurs moyenne non adaptés. Cette combinaison linéaire est estimée au maximum de vraisemblance des données d'adaptation. Soit G_i un noeud de l'arbre contenant les M_i gaussiennes $\{g_{(i,1)}, g_{(i,2)}, \cdots, g_{(i,M_i)}\}$ et $O = (o_1, o_2, \cdots, o_T)$ l'ensemble des T vecteurs d'observations issus des données d'adaptation. Soient $(\mu_{(i,m)}, \Sigma_{(i,m)})$ respectivement le vecteur moyenne de dimension n et la matrice de covariance de dimension $n \times n$ de la gaussienne m du noeud G_i. *SMLLR* modifie linéairement chaque vecteur moyenne $\mu_{(i,m)}$ du noeud G_i, pour $m = 1, 2, \cdots, M_i$, en lui appliquant une matrice $\hat{W}_i$ de dimension $n \times (n+1)$ selon l'équation :

$$\hat{\mu}_{(i,m)} \quad = \quad \hat{W}_i \; \xi_{(i,m)} \tag{1}$$

où $\xi_{(i,m)}$ est le vecteur étendu du vecteur moyenne $\mu_{(i,m)}$ tel que $\xi_{(i,m)} = [1 \; \mu^{'}_{(i,m)}]^{'}$ et $\hat{\mu}_{(i,m)}$ le vecteur moyenne adapté.
Chaque $\hat{W}_i$ est estimée selon l'équation :

$$\sum_{m=1}^{M_i} \sum_{t=1}^{T} \gamma_t(g_{(i,m)}) \; \Sigma^{-1}_{(i,m)} \; o_t \; \xi^{'}_{(i,m)} \quad =$$

$$\sum_{m=1}^{M_i} \sum_{t=1}^{T} \gamma_t(g_{(i,m)}) \; \Sigma^{-1}_{(i,m)} \; \hat{W}_i \; \xi_{(i,m)} \; \xi^{'}_{(i,m)} \tag{2}$$

où $\gamma_t(g_{(i,m)})$ est la probabilité d'occuper la gaussienne m du noeud G_i à l'instant t en sachant que la séquence O a été observée. En supposant que toutes les matrices de covariance $\Sigma_{(i,m)}$ sont diagonales, le calcul de la matrice $\hat{W}_i$ s'effectue ligne par ligne en résolvant les n systèmes d'équations linéaires. Chaque système est constitué de $n+1$ équations à $n+1$ inconnues.
Chaque noeud G_i auquel est associé une matrice de transformation $\hat{W}_i$ est appelé une classe de régression. La technique *SMLLR* que nous avons expérimenté détermine ces classes dynamiquement en fonction de la quantité de données d'adaptation disponible. Pour déterminer l'ensemble des classes de régression, l'arbre des gaussiennes est tout d'abord construit selon la méthode exposée au paragraphe précédent. Tous les noeuds feuilles sont ensuite retenus comme classes de régression. Pour chaque noeud feuille qui disposent d'au moins $MinOb$ observations, la transformation qui y est associée est estimée en utilisant l'ensemble des gaussiennes du noeud. Par contre, pour chaque noeud feuille ne disposant pas d'un nombre suffisant d'observations, la transformation associée est estimée en utilisant les gaussiennes du noeud ascendant le plus proche disposant d'un nombre suffisant d'observations. Ce processus de constitution des classes de régression permet de mettre à jour les paramètres des gaussiennes en utilisant des matrices de transformation estimées de manière fiable.

2.3. SMAP

Comme pour *SMLLR*, nous considérons que seule l'adaptation des moyennes des gaussiennes est réalisée. On suppose tout d'abord l'arbre des gaussiennes construit. Soit $G_i = \{g_{(i,1)}, g_{(i,2)}, \cdots, g_{(i,M_i)}\}$ un noeud de cet arbre contenant M_i gaussiennes et $O = (o_1, o_2, \cdots, o_T)$ l'ensemble des T vecteurs d'observations issus des données d'adaptation.
SMAP [12] transforme chaque vecteur o_t en un vecteur y_{mt} pour chaque gaussienne m selon l'équation :

$$y_{mt} \quad = \quad \Sigma_m^{(-1/2)} \; (o_t - \mu_m) \tag{3}$$

pour $t = 1, \cdots, T$ et $m = 1, \cdots, M$.
μ_m et Σ_m sont respectivement le vecteur moyenne et la matrice de covariance de la gaussienne m, M est le nombre total de gaussiennes. *SMAP* suppose que $Y_i = \{Y_{(i,m)}\}_{m \in G_i}$, telle que $Y_{(i,m)} = \{y_{m1}, y_{m2}, \cdots, y_{mT}\}$ peut être modélisée par une distribution normale $\mathcal{N}(\nu^{(i)}, \eta^{(i)})$. Les paramètres $\nu^{(i)}$ et $\eta^{(i)}$ permettent de réestimer respectivement le vecteur moyenne et la matrice de covariance des gaussiennes du noeud G_i. Comme nous avons choisi de ne modifier que les moyennes des gaussiennes des modèles acoustiques [4], seul le paramètre $\nu^{(i)}$ est estimé. $\nu^{(i)}$ est estimé au maximum a posteriori des données d'adaptation. Il est estimé pour chaque noeud de l'arbre des gaussiennes. Au niveau des feuilles, chaque noeud G_i représente une seule gaussienne. Le paramètre $\nu^{(i)}$ permet de mettre à jour la moyenne des gaussiennes du noeud G_i selon l'équation :

$$\hat{\mu}_m \quad = \quad \mu_m + (\Sigma_m)^{(1/2)} \; \hat{\nu}^{(i)} \tag{4}$$

pour toutes les gaussiennes m telles que $m \in G_i$.
Soit G_f un noeud et G_p son noeud père. *SMAP* émet l'hypothèse que l'estimée $\hat{\nu}^{(f)}$ de $\nu^{(f)}$ au noeud G_f est obtenue en utilisant l'estimée au maximum de vraisemblance $\tilde{\nu}^{(f)}$ de $\nu^{(f)}$ et l'estimée a posteriori du noeud père $\hat{\nu}^{(p)}$. L'estimée $\hat{\nu}^{(f)}$ s'obtient selon l'équation :

$$\hat{\nu}^{(f)} \quad = \quad \frac{\Gamma^{(f)} \; \tilde{\nu}^{(f)} \; + \; \tau^{(f)} \; \hat{\nu}^{(p)}}{\Gamma^{(f)} + \tau^{(f)}} \tag{5}$$

où $\Gamma^{(f)} = \sum_{m=1}^{M_f} \; \sum_{t=1}^{T} \; \gamma_t(g_{(f,m)})$.
$\tau^{(f)}$ est un hyperparamètre qui détermine l'influence de

[4] Ce choix est motivé par la volonté de fournir un cadre commun pour la comparaison de *SMLLR* et de *SMAP*.

$\hat{\nu}^{(p)}$ sur $\tilde{\nu}^{(f)}$. Il est soit choisi empiriquement et demeure constant pour tous les noeuds G_f [12], soit il est estimé au maximum de vraisemblance [6] ou au maximum a posteriori [7].
L'estimée au maximum de vraisemblance $\tilde{\nu}^{(f)}$ au noeud G_f s'obtient selon l'équation :

$$\tilde{\nu}^{(f)} = \frac{\sum_{m=1}^{M_f} \sum_{t=1}^{T} \gamma_t(g_{(f,m)}) \, \Sigma_{(f,m)}^{(-1/2)} \left(o_t - \mu_{(f,m)}\right)}{\sum_{m=1}^{M_f} \sum_{t=1}^{T} \gamma_t(g_{(f,m)})}$$

3. VALIDATION EXPÉRIMENTALE

3.1. Conditions expérimentales

SMLLR et *SMAP* ont été implantées dans le système de reconnaissance ESPERE [3] et testées sur le corpus *RM*. Le système ESPERE est une boîte à outils pour la reconnaissance de la parole basée sur les HMMs du premier ordre. Le corpus *RM* a été utilisé de la façon suivante :

Apprentissage du système indépendant du locuteur : partie indépendante du locuteur (RM1). Elle rassemble 3360 phrases prononcées par 80 locuteurs natifs américains, chacun ayant fournit 42 phrases.

Adaptation et tests : partie dépendante du locuteur (RM2). Elle regroupe quatre locuteurs. Chacun d'eux a prononcé 600 phrases d'apprentissage (utilisées uniquement pour l'adaptation) et 120 phrases de test (utilisées en phase de tests).

Le système indépendant du locuteur comprend 45 HMMs à 3 états et un HMM à un état pour le silence. La densité d'observation de chaque état d'un HMM est modélisée par un mélange de 8 gaussiennes. De chaque trame de parole est extrait un vecteur d'observation composé de 35 coefficients cepstraux [5]. L'apprentissage des modèles acoustiques du système indépendant du locuteur a été réalisé avec l'algorithme de Baum-Welch, en 20 itérations. Les tests de reconnaissance ont été réalisés en utilisant la grammaire standard *word-pair* de *RM*. Le système indépendant du locuteur a été adapté pour chaque locuteur en utilisant un nombre différent de phrases d'adaptation. L'adaptation a été réalisée en mode par lot supervisé, en mode par lot non supervisé et en mode incrémental non supervisé.
La méthode utilisée pour construire l'arbre des gaussiennes utilisé par *SMLLR* et *SMAP* est la méthode descendante *LBG* combinée à la méthode des *K-Means* pour affiner les classes. L'arbre construit est un arbre binaire. Nous avons utilisé la distance de Mahalanobis comme mesure de distance entre un centre de gravité et une gaussienne.
Le mode incrémental utilisé est celui proposé dans [2]. La méthode consiste à récolter les statistiques suffisantes pour chaque gaussienne à l'aide de la procédure *forward-backward* en utilisant le système précédemment adapté et le bloc de phrases courant. Ces statistiques sont ajoutées aux statistiques éventuellement récoltées pour les blocs de phrases précédentes et ce cumul des statistiques est utilisé pour l'estimation des paramètres.
Le système ESPERE a été testé en mode dépendant du locuteur, indépendant du locuteur et en mode adapté au locuteur. Tous les résultats qui suivent représentent les performances moyennes sur les quatre locuteurs. Les performances moyennes du système dépendant du locuteur sont de $94,9\%$, celles du système indépendant du locuteur de $88,6\%$, en prenant un intervalle de confiance de $\pm 1\%$, avec un risque de 5%.

[5] Les 11 cepstres $C1$ à $C11$, les 12 dérivées premières et les 12 dérivées secondes des cepstres $C0$ à $C11$.

3.2. Résultats

SMLLR Nous avons utilisé un arbre de profondeur 6, avec un nombre minimum d'observations de 1000 pour la constitution des classes de régression. Ce paramétrage nous a permis d'obtenir les meilleurs résultats.
Le tableau 1 montre que les performances obtenues à l'issue d'une adaptation par lot en utilisant un arbre sont meilleures qu'en utilisant des classes statiques définies à l'aide de connaissances phonétiques, quelque soit le nombre de phrases d'adaptation disponibles. L'utilisation de classes phonétiques impose en effet d'estimer les paramètres d'un certain nombre de transformations, qui n'est pas forcément le plus adéquat par rapport à la quantité de données d'adaptation disponible.

	1	5	10	100
Arbre prof. 6	88.6	90.3	90.7	91.5
1 classe phonétique	87.9	90.3	90.4	90.7
3 classes phonétiques	42	88.4	89.4	91
6 classes phonétiques	-	80.9	85.1	90.7
10 classes phonétiques	-	54.5	79.9	91

Table 1 – Adaptation *MLLR* avec classes phonétiques (Taux de reconnaissance Mot)

Les performances en mode supervisé et en mode non supervisé (figure 1) sont assez similaires. Ce phénomène peut s'expliquer par le fait que le système indépendant du locuteur obtient des performances raisonnables. Les erreurs de reconnaissance qui se propagent dans le processus d'adaptation restent donc limitées. Le gain en performances devient significatif à partir de trois phrases d'adaptation.

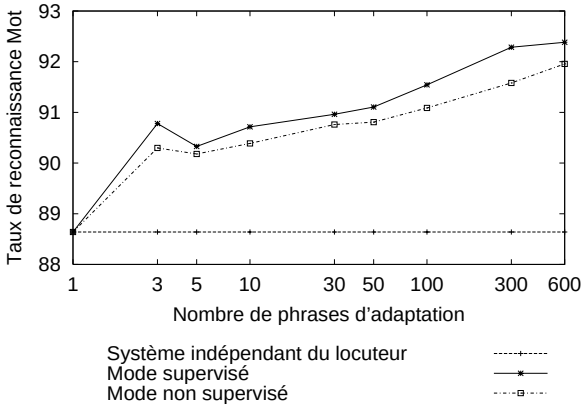

Figure 1 – Adaptation *SMLLR* par lot

En mode incrémental non supervisé (figure 2), l'utilisation de blocs d'une seule phrase permet au SRAP de s'adapter très rapidement aux nouvelles conditions acoustiques. Comme *SMLLR* est paramétré pour qu'un nombre suffisant d'observations soit requis pour obtenir une estimation fiable des paramètres, des blocs d'une seule phrase peuvent être utilisés sans risque de dégradation précoce des performances.

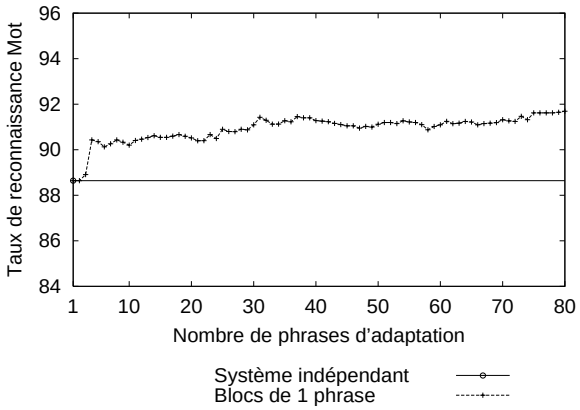

Figure 2 – Adaptation *SMLLR* incrémentale non supervisée

SMAP Pour pouvoir comparer *SMAP* à *SMLLR*, nous avons utilisé le même arbre de gaussiennes que *SMLLR* pour *SMAP*. Le choix de l'hyperparamètre $\tau^{(i)}$ associé à un noeud G_i (équation 5) a été déterminé expérimentalement. $\tau^{(i)}$ reste constant pour tous les noeuds G_i et dépend de la quantité totale de données d'adaptation disponible. Plus la quantité de données d'adaptation est faible, plus τ est grand, permettant de prendre plus en compte les probabilités a priori dans l'estimation des paramètres. Plus la quantité de données est grande, plus τ devient petit.
Comme *SMLLR* dans le cas d'une adaptation par lot, *SMAP* donne des performances assez similaires en mode supervisé et en mode non supervisé (figure 3). *SMAP* permet d'obtenir une amélioration des performances même avec peu de données d'adaptation, à l'inverse de *SMLLR*. Ceci est dû à l'utilisation des probabilités a priori sur les paramètres dans le cas où l'estimation des paramètres est limitée par la quantité de données d'adaptation disponible. Pour *SMAP*, l'amélioration des performances est significative à partir de dix phrases d'adaptation.

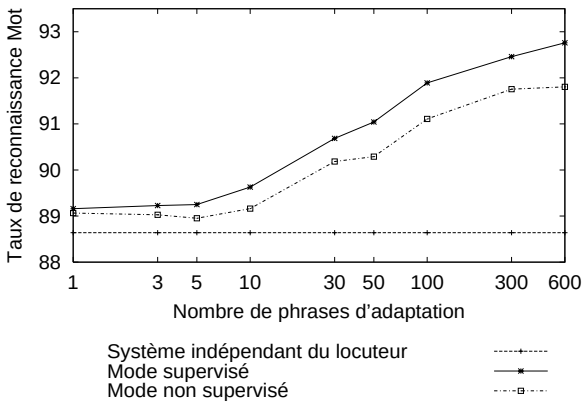

Figure 3 – Adaptation SMAP par lot

Dans le cas d'une adaptation incrémentale non supervisée (figure 4), *SMAP* se révèle moins performante que *SMLLR*. Nous n'avons pas fait figurer les performances obtenues pour des blocs d'une phrase. Ces performances se sont en effet montrées désastreuses pour deux locuteurs. Nous pensons que la technique incrémentale que nous avons utilisé pour *SMAP* n'est pas compatible avec le processus d'estimation de cette méthode. Nous effectuons actuellement des investigations pour résoudre ce problème.

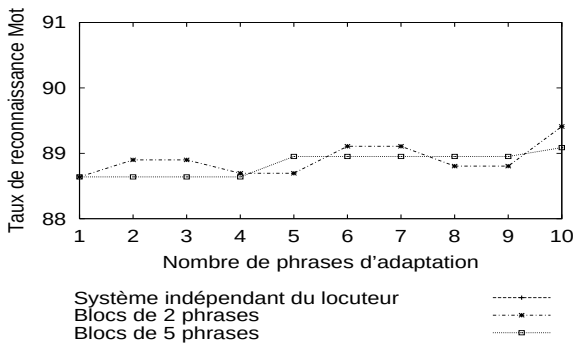

Figure 4 – Adaptation SMAP incrémental non supervisée

4. CONCLUSION

Nous avons présenté et évalué expérimentalement les techniques d'adaptation *SMLLR* et *SMAP* pour le moteur de reconnaissance ESPERE. Ces deux techniques permettent d'améliorer de manière significative les performances d'un système indépendant du locuteur en mode par lot. *SMLLR* est plus efficace, pour la technique incrémentale choisie, pour une adaptation incrémentale non supervisée.
Nous envisageons par la suite d'expérimenter *SMLLR* en utilisant des matrices de transformation blocs diagonales. Nous prévoyons également d'étudier plusieurs méthodes d'élagage des noeuds d'un arbre de gaussiennes, afin de choisir celle qui permet d'obtenir l'arbre prenant le mieux en compte les données d'adaptation. Nous nous orientons actuellement vers des techniques d'adaptation incrémentales non supervisées.

BIBLIOGRAPHIE

[1] M. Afify and O. Siohan. Constrained Maximum Likelihood Linear Regression for speaker adaptation. *ICSLP*, pages 861–864, 2000.

[2] V.V. Digalakis. Online adaptation Hidden Markov Models using incremental estimation algorithms. *TSAP*, 7(3) :253–261, 1999.

[3] D. Fohr, O. Mella, and C. Antoine. The automatic speech recognition engine ESPERE : experiments on telephone speech. *ICSLP*, pages 246–249, 2000.

[4] M.J.F. Gales and P.C. Woodland. Mean and variance adaptation within the MLLR framework. *Computer Speech and Language*, 10 :249–264, 1996.

[5] J.-L. Gauvain and C.-H. Lee. Maximum A Posteriori estimation for multivariate gaussian mixture observations of Markov chains. *TSAP*, 2(2) :291–298, 1994.

[6] Q. Huo and C.-H. Lee. On-line adaptive learning of the Continuous Density Hidden Markov Model based on approximate recursive Bayes estimate. *TSAP*, 5(2) :161–171, 1997.

[7] Q. Huo and C.-H. Lee. On-line adaptive learning of the correlated Continuous Density Hidden Markov Models for speech recognition. *TSAP*, 6(4) :386–397, 1998.

[8] C.J. Leggetter and P.C. Woodland. Flexible speaker adaptation using Maximum Likelihood Linear Regression. *Eurospeech'95*, pages 1155–1158, 1995.

[9] C.J. Leggetter and P.C. Woodland. Maximum Likelihood Linear Regression for speaker adaptation of Continous Density Hidden Markov Models. *Computer Speech and Language*, 9 :171–185, 1995.

[10] T.A. Myrvoll, O. Siohan, C.-H. Lee, and W. Chou. Structural Maximum A Posteriori Linear Regression for unsupervised speaker adaptation. *ICSLP*, pages 540–543, 2000.

[11] K. Shinoda and C.-H. Lee. Unsupervised adaptation using structural Bayes approach. *ICASSP*, pages 793–796, 1998.

[12] K. Shinoda and C.-H. Lee. A structural Bayes approach to speaker adaptation. *TSAP*, 9(3) :276–287, 2001.