HAL open science

Adaptation MLLR pour des HMMs

Fabrice Lauri, Irina Illina, Dominique Fohr

To cite this version:

Fabrice Lauri, Irina Illina, Dominique Fohr. Adaptation MLLR pour des HMMs. Quatrièmes Rencontres Jeunes Chercheurs en Parole - RJC'2001, 2001, Mons, Belgique, pp.90-93. inria-00147543

HAL Id: inria-00147543

https://inria.hal.science/inria-00147543

Submitted on 18 May 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Adaptation MLLR pour des HMMs

Fabrice Lauri, Irina Illina, Dominique Fohr

LORIA
BP 239 - 54506 Vandoeuvre-Lès-Nancy
Tél.: 03 83 59 20 55
Mél: lauri@loria.fr

Résumé

Nous présentons dans cet article la technique d'adaptation *Maximum Likelihood Linear Regression* (MLLR). MLLR permet d'adapter les paramètres des modèles acoustiques d'un système indépendant du locuteur afin d'améliorer la reconnaissance pour un nouvel environnement de test. Nous avons expérimenté MLLR à des HMMs du moteur ESPERE appris sur le corpus de parole *Resource Management* (RM). Les résultats montrent une réduction de 12,6 % d'erreurs en moyenne par rapport au système indépendant du locuteur, en utilisant 30 phrases d'adaptation.

1. Introduction

Actuellement, les systèmes de reconnaissance sont utilisés en mode soit indépendant du locuteur, soit dépendant du locuteur. Les premiers peuvent s'utiliser pratiquement immédiatement, avec cependant un taux de reconnaissace de deux à trois fois inférieur aux systèmes dépendants du locuteur [Lee91]. Ces derniers nécessitent quant à eux de la part du locuteur une prise en main fastidieuse. Le locuteur doit en effet fournir un apport important en données de parole indispensable à l'apprentissage robuste du système. En outre, lorsque ces deux types de systèmes sont utilisés dans un environnement différent de celui ayant servi à l'apprentissage (environnement bruité ou nouveau locuteur), leurs performances chutent. Les techniques d'adaptation permettent alors de réduire les différences entre les conditions d'apprentissage et les conditions de tests, ce qui améliore les taux de reconnaissance du système dans l'environnement de test.
Cet article présente la méthode d'adaptation MLLR, proposée par *Leggetter et Woodland* dans [Leg95]. MLLR utilise une transformation pour modifier les paramètres des modèles acoustiques. Elle est désormais l'une des méthodes d'adaptation les plus utilisées et l'une des plus déclinées dans le domaine de l'adaptation des modèles acoustiques d'un système indépendant du locuteur.
Plusieurs améliorations de MLLR ont été proposées récemment. Dans le cas de données d'adaptation dispersées et en faible quantité, Afify et Siohan proposent d'utiliser comme transformation une transformation diagonale d'une certaine largeur de bande [Afi00]. Dans [Sio00] et [SiM00], Siohan estime la transformation au *Maximum a Posteriori* (MAP). L'estimation au MAP de la transformation ou sa restriction à une structure donnée assurent une structure probabiliste stable des modèles acoustiques lorsque peu de données d'adaptation sont disponibles.
Dans [Boc99] et [Doh00], les auteurs modélisent des dépendances entre les paramètres des modèles acoustiques qui ont des propriétés différentes [1] ([Boc99], [Doh00]). D'une part, cela permet d'affiner les paramètres des modèles acoustiques pour lesquelles peu de données d'adaptation ont été utilisées. D'autre part, les paramètres des modèles n'ayant pas été adaptés, faute de données d'adaptation disponibles, peuvent être estimés.

La suite de cet article est organisée de la façon suivante. Dans un premier temps, nous exposons la méthode d'adaptation MLLR. Nous indiquons ensuite les expériences réalisées avec cette technique. Nous concluons enfin sur les résultats et nos perspectives de recherche envisagées.

2. *Maximum Likelihood Linear Regression*

MLLR est une technique d'adaptation qui transforme les paramètres des modèles acoustiques d'un système indépendant du locuteur. Cela permet de réduire les différences acoustiques entre les conditions d'apprentissage et les conditions de tests et améliore ainsi le taux de reconnaissance du système pour un nouveau locuteur ou un nouvel environnement.

2.1. *Principe général*

Dans le cas le plus général, on cherche à estimer une transformation W qui modélise les différences supposées linéaires entre les conditions d'apprentissage et les conditions de tests. Le système adapté est obtenu en appliquant cette transformation linéaire W à un ensemble des paramètres des modèles issus de l'apprentissage.

MLLR peut être appliqué à des modèles acoustiques représentés par des HMMs à densité de probabilité continue. Une densité est modélisée par un mélange de gaussiennes. La densité de probabilité de l'observation o à l'état i d'un HMM est donnée par :

$$b_i(o) = \sum_{m=1}^{M_i} c_{i,m}\, b_{i,m}(o)$$

[1] Par exemple, une relation de dépendance entre les paramètres des modèles acoustiques appartenant à la famille des occlusives et les paramètres des modèles acoustiques appartenant aux fricatives pourra être modélisée.

où
- M_i est le nombre de lois gaussiennes à l'état i,
- $b_{i,m}(o)$ est la m-ème densité gaussienne associée à l'état i telle que

$$b_{i,m}(o) = \mathcal{N}(o, \mu_{i,m}, C_{i,m}) =$$

$$\frac{1}{(2\pi)^{n/2}|C_{i,m}|^{1/2}} exp\left[-\frac{1}{2}(o-\mu_{i,m})^t C_{i,m}^{-1}(o-\mu_{i,m})\right]$$

- $\mu_{i,m}$ est le vecteur moyenne de dimension $n \times 1$ de la gaussienne m associé à l'état i,
- $C_{i,m}$ est la matrice de covariance de dimension $n \times n$ de la gaussienne m associé à l'état i.

Nous supposons que les différences acoustiques entre les conditions d'apprentissage et les conditions de tests sont principalement caractérisées par les moyennes des modèles. Dans ce cas, chaque moyenne $\hat{\mu}_{i,m}$ du système adapté est obtenue par transformation linéaire de la moyenne $\mu_{i,m}$ du modèle initial :

$$\hat{\mu}_{i,m} = W\ \xi_{i,m}$$

où $\xi_{i,m} = (\mu_{i,m}^t\ \ 1)^t$ est le vecteur étendu de la moyenne de la gaussienne m à l'état i et W est la matrice de transformation de dimension $n \times (n+1)$.

Pour obtenir une adaptation plus robuste des paramètres, plusieurs transformations peuvent être utilisées. Une transformation est alors appliquée à un ensemble de gaussiennes, regroupées selon un critère donné. Cet ensemble de gaussiennes est appelé une *classe de régression*. Une classe de régression lie ainsi plusieurs gaussiennes à une même transformation.

La figure 1 présente le processus d'adaptation MLLR de quatre modèles acoustiques (HMMs à trois états). L'adaptation utilise deux classes de régression, avec deux modèles acoustiques par classe.

Nous abordons maintenant le problème de l'estimation de la transformation W.

2.2. Estimation de W

Nous considérons le cas d'un HMM constitué de I états à densité continue modélisée par un mélange de M gaussiennes [2]. Une seule matrice de régression W sera utilisée pour modifier l'ensemble des moyennes des modèles du système.

Dans le système adapté, les moyennes sont obtenues par transformation linéaire des moyennes initiales. La densité de probabilité $b_{i,m}(o)$ devient alors :

$$b_{i,m}(o) =$$

$$\frac{1}{(2\pi)^{n/2}|C_{i,m}|^{1/2}} exp[-\frac{1}{2}(o-W\xi_{i,m})^t C_{i,m}^{-1}(o-W\xi_{i,m})]$$

L'ensemble des paramètres λ des modèles acoustiques devient $\lambda = (\pi, A, B, W)$ [3]. W est estimée au maximum de vraisemblance des données d'adaptation $O = (o_1, o_2, ..., o_T)$, soit :

$$\hat{W} = \underset{W}{\text{argmax}}\ p(O/\lambda) \qquad (1)$$

Or

$$\begin{aligned} p(O/\lambda) &= \sum_{S\in\Phi} p(O,S/\lambda) \\ &= \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} b_{s_t}(o_t) \\ &= \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} \left[\sum_{m=1}^{M} c_{s_t,m} b_{s_t,m}(o_t)\right] \end{aligned}$$

avec $p(O,S/\lambda)$ la probabilité de générer la séquence d'observations O en passant par le chemin $S = (s_1, s_2, ..., s_T)$ et Φ l'ensemble des séquences d'états S de longueur T.

Soit $\Psi = \{1, 2, ..., M\}^T$ l'ensemble des séquences de mixtures $K = (k_1, k_2, ..., k_T)$ de longueur T et $p(O,S,K/\lambda)$ la probabilité jointe de O, S et K. Alors, $p(O,S/\lambda)$ peut s'écrire comme la probabilité marginale de $p(O,S,K)$ sur Ψ telle que :

$$\begin{aligned} p(O,S/\lambda) &= \sum_{K\in\Psi} p(O,S,K/\lambda) \\ &= \sum_{K\in\Psi} \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} c_{s_t,k_t} b_{s_t,k_t}(o_t) \end{aligned}$$

L'équation 1 se réécrit alors comme :

$$\hat{W} = \underset{W}{\text{argmax}} \sum_{S\in\Phi} \sum_{K\in\Psi} p(O,S,K/\lambda) \qquad (2)$$

L'algorithme EM permet de trouver la matrice $\hat{W}$ qui maximise $p(O/\lambda)$. Soit la fonction auxiliaire Q telle que :

$$Q(\lambda, \hat{\lambda}) = \frac{1}{p(O/\lambda)} \sum_{S\in\Phi} \sum_{K\in\Psi} p(O,S,K/\lambda)\ log\ p(O,S,K/\hat{\lambda})$$

où $\hat{\lambda}$ est l'ensemble des paramètres à estimer et λ l'ensemble des paramètres actuels. En maximisant la fonction Q par rapport à $\hat{\lambda}$, en remplaçant λ par $\hat{\lambda}$ et en itérant ce processus jusqu'à ce que la différence $Q(\lambda, \hat{\lambda}) - Q(\lambda, \lambda)$ converge, on est assuré que le paramètre W est celui qui maximise localement $p(O/\lambda)$.

Comme seule W est estimée, seule la densité de probabilité $b_{s_t,k_t}(o_t)$ est affectée, d'où :

$$Q(\lambda, \hat{\lambda}) = cst +$$

$$\frac{1}{p(O/\lambda)} \sum_{S\in\Phi} \sum_{K\in\Psi} \sum_{t=1}^{T} p(O,S,K/\lambda)\ log\ \hat{b}_{s_t,k_t}(o_t) \qquad (3)$$

En posant :

$$\begin{aligned} \sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_t(i,m) &= \sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{t=1}^{T} p(s_t = i, k_t = m/O, \lambda) \\ &= \sum_{S\in\Phi} \sum_{K\in\Psi} p(S,K/O,\lambda) \end{aligned}$$

[2] Tous les états sont donc supposés avoir le même nombre de gaussiennes ($M_i = M\ \forall i$).

[3] π est le vecteur des probabilités initiales, A est la matrice des probabilités de transition et B est l'ensemble des densités de probabilité d'observation

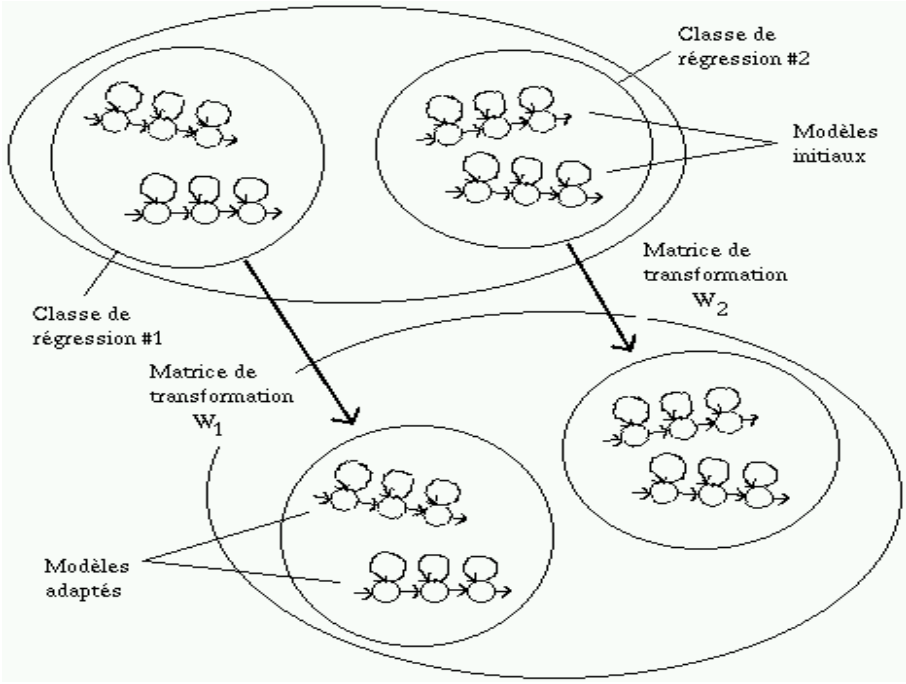

FIG. 1 – MLLR avec deux classes de régression

où $\gamma_t(i,m)$ est la probabilité *a posteriori* de se trouver dans la gaussienne m de l'état i à l'instant t en sachant que la séquence d'observations O a été généré. Chaque $\gamma_t(i,m)$, pour $1 \leq i \leq I$, $1 \leq m \leq M$ et $1 \leq t \leq T$ est obtenue à partir du corpus d'adaptation à l'aide de l'algorithme de Baum-Welch.
L'équation (3) devient alors :

$$Q(\lambda, \hat{\lambda}) = cst + \sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_t(i,m) \, log \, \hat{b}_{i,m}(o_t)$$

$$Q(\lambda, \hat{\lambda}) = cst + \sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_t(i,m)$$

$$\left[-\frac{1}{2} \left[n \, log \, 2\pi + log \, |C_{i,m}| + h(o_t, i, m) \right] \right]$$

avec

$$h(o_t, i, m) = (o_t - \hat{W} \xi_{i,m})^t C_{i,m}^{-1} (o_t - \hat{W} \xi_{i,m})$$

Trouver W qui maximise la fonction $Q(\lambda, \hat{\lambda})$ revient à dériver Q par rapport à W puis à annuler la dérivée, soit :

$$\frac{\partial}{\partial \hat{W}} Q(\lambda / \hat{\lambda}) = \sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_t(i,m) C_{i,m}^{-1} (o_t - \hat{W} \xi_{i,m}) \xi_{i,m}^t$$

$$\frac{\partial}{\partial \hat{W}} Q(\lambda / \hat{\lambda}) \equiv 0$$

ce qui donne :

$$\sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_t(i,m) C_{i,m}^{-1} o_t \xi_{i,m}^t =$$

$$\sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_t(i,m) C_{i,m}^{-1} \hat{W} \xi_{i,m} \xi_{i,m}^t$$

Le calcul de $\hat{W}$ s'effectue ligne par ligne en résolvant les n systèmes de $n+1$ équations linéaires, dans le cas où $C_{i,m}$ est diagonale.

3. Validation expérimentale

3.1. Description des expériences

L'approche d'adaptation MLLR a été testée sur le système de reconnaissance ESPERE [Foh00]. Ce système est basé sur les HMMs du premier ordre. Le corpus d'apprentissage RM a été utilisé pour générer un système indépendant du locuteur et réaliser les tests d'adaptation. Nous avons utilisé l'ensemble RM de la façon suivante :

apprentissage : 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 (utilisé uniquement pour l'adaptation) et 120 phrases de test (utilisé en phase de tests).

Le système indépendant du locuteur comprend 46 HMMs à 3 états dont 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 [4]. L'apprentissage des modèles du système indépendant du locuteur a été réalisé avec l'algorithme de Baum-Welch, en 20 itérations. Le système indépendant du locuteur a été adapté selon chaque locuteur avec un nombre de phrases d'adaptation donné. Les tests de reconnaissance ont été réalisés en utilisant la grammaire standard *word-pair* de RM. Le système ESPERE a été testé en mode dépendant du locuteur, indépendant du locuteur et en mode adapté au locuteur.

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

3.2. Résultats

Tous les tests d'adaptation ont été réalisés de manière supervisée : la transcription phonétique de chaque phrase d'adaptation a toujours été indiquée. L'adaptation s'est réalisée avec une seule itération. En outre, une seule classe de régression a été utilisée lors de l'adaptation. La figure 2 présente le taux d'erreur moyen de reconnaissance des quatre systèmes adaptés avec MLLR (un système par locuteur). Lorsque peu de données d'adaptation sont disponibles, les erreurs d'estimation de la matrice de transformation sont importantes. Ces erreurs engendrent une hausse du taux d'erreur des mots du système adapté par rapport à celui du système indépendant du locuteur. Plus la quantité de données d'adaptation devient importante, plus l'estimation de la matrice de régression est précise. En utilisant 30 phrases d'adaptation, le taux d'erreur peut être réduit de 12,6 % en moyenne par rapport à celui du système indépendant du locuteur. En utilisant 200 phrases, ce taux est réduit de 15 %.

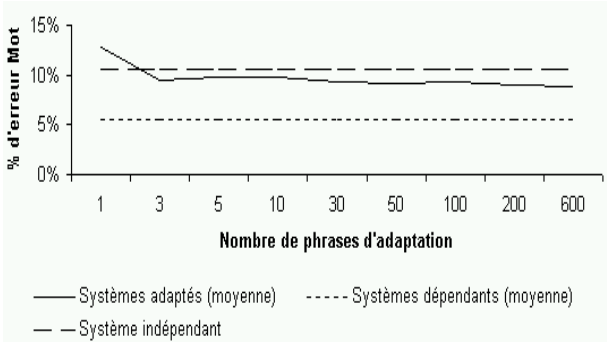

FIG. 2 – Performances de MLLR avec une classe de régression

La figure 3 indique les performances individuels des systèmes adaptés à chaque locuteur. On observe une baisse du taux d'erreur des mots pour les systèmes adaptés respectivement pour les locuteurs 1, 3 et 4. En revanche, le taux d'erreur du système adapté pour le locuteur 2 a augmenté par rapport à celui du système indépendant du locuteur. Nous expliquons ce phénomène par une trop grande différence acoustique entre les données utilisées lors de l'adaptation et celles utilisées lors des tests de reconnaissance. Enfin, c'est sur le locuteur 4 que MLLR a donné les baisses du taux d'erreur. Le taux d'erreur des mots du système adapté pour le locuteur 4 est en effet réduit de 50 % par rapport à celui du système indépendant du locuteur.

4. Conclusion

Nous avons présenté et testé la technique d'adaptation MLLR pour le moteur de reconnaissance ESPERE. Cette technique permet actuellement d'améliorer de manière significative le taux de reconnaissance pour un locuteur, à condition de disposer d'un nombre suffisant de données d'adaptation. Des tests d'adaptation avec plusieurs classes de régression sont en cours. Nous envisageons par la suite de décliner cette technique d'adaptation vers une technique d'adaptation incrémentale, qui nécessiterait moins de temps de calcul et permettrait d'améliorer le taux de reconnaissance d'un locuteur au fur et à mesure qu'il parle.

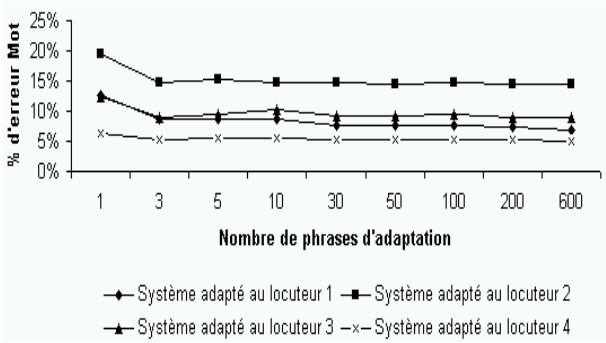

FIG. 3 – Performances MLLR pour plusieurs locuteurs

Références

[Lee91] Lee C.-H, Lin C.-H., Juang B.-H. *A study on speaker adaptation of the parameters of continuous density hidden markov models.* IEEE Transactions on Signal Processing, Vol. 39, pp. 806-814, 1991.

[Leg95] Leggetter C. J., Woodland P. C. *Maximum likelihood linear regression for speaker adaptation of continous density hidden Markov models.* Computer Speech and Language, Vol. 9, pp. 171-185, 1995.

[Sio00] Siohan O., Chesta C., Lee C.-H. *Joint Maximum A Posteriori estimation of transformation and Hidden Markov Models parameters.* Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Istanbul, Turkey, pp. 965-968, 2000.

[Boc99] Bocchieri E., Digalakis V., Corduneanu A., Boulis C. *Correlation modeling of MLLR transform biases for rapid HMM adaptation to new speakers.* IEEE International Conference On Acoustics, Speech and Signal Processing, vol. 2, pp. 773-776, 1999.

[Doh00] Doh S.-J., Stern R.M. *Inter-Class MLLR for speaker adaptation.* IEEE Conf. on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, pp. 1543-1546, Juin 2000.

[Afi00] Afify M., Siohan O. *Constrained Maximum Likelihood Linear Regression for speaker adaptation.* Proc. Int. Conf. on Spoken Language Processing, Beijing, China, pp. 861-864, 2000.

[SiM00] Siohan O., Myrvoll T.A., Lee C.-H. *Structural Maximum A Posteriori Linear Regression for fast HMM adaptation.* Workshop on Automatic Speech Recognition : Challenges for the new Millenium, Paris, France, pp. 120-127, Septembre 2000.

[Foh00] Fohr D., Mella O., Antoine C. *The automatic speech recognition engine ESPERE : experiments on telephone speech.* ICSLP, Pékin, Chine. 2000.