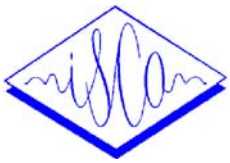

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Acoustic model adaptation based on coarse/fine training of transfer vectors and its application to a speaker adaptation task

Shinji Watanabe and Atsushi Nakamura

NTT Communication Science Laboratories, NTT Corporation
{watanabe, ats}@cslab.kecl.ntt.co.jp

Abstract

In this paper, we propose a novel adaptation technique based on coarse/fine training of transfer vectors. We focus on transfer vector estimation of a Gaussian mean from an initial model to an adapted model. The transfer vector is decomposed into a direction vector and a scaling factor. By using tied-Gaussian class (coarse class) estimation for the direction vector, and by using individual Gaussian class (fine class) estimation for the scaling factor, we can obtain accurate transfer vectors with a small number of parameters. Simple training algorithms for transfer vector estimation are analytically derived using the variational Bayes, maximum a posteriori (MAP) and maximum likelihood methods. Speaker adaptation experiments show that our proposals clearly improve speech recognition performance for any amount of adaptation data, compared with conventional MAP adaptation.

1. Introduction

Acoustic models for speech recognition are obtained by training based on speech data. Since speech varies due to such factors as speaker, speaking style, noise, etc, it is unrealistic to collect a complete set of data that contains every possible speech variation. Consequently, speech recognizers often encounter speech inputs that do not match the acoustic models, resulting in degraded performance. In order to adjust acoustic models so that they quickly match every type of speech, it is necessary to adapt the model using only a small amount of data.

Model adaptation techniques are promising solutions for such quick adjustments of acoustic models. Since conventional Maximum Likelihood (ML) based estimation often causes over-training when the amount of data is insufficient, several proposals have been made for adaptation techniques that avoid over-training [1–6]. Bayesian approaches avoid over-training by utilizing Bayesian priors where the estimation parameter belongs to individual Gaussians [2]. Transformation estimation approaches avoid over-training by estimating the transformation of model parameters using tied-Gaussian classes [3]. The performance of transformation approaches is usually better than Bayesian approaches for small amounts of data because they require fewer parameters to be estimated, due to the use of tied-Gaussian class (coarse class) estimation. However, the performance of transformation approaches with a large amount of data is worse than Bayesian approaches due to the use of individual Gaussian class (fine class) estimation in Bayesian approaches. Therefore, a new adaptation technique which optimally combines both coarse and fine estimation is desired to improve adapted models for any amount of data.

In this paper, we propose an adaptation technique that uses both coarse and fine estimation. We focus on transfer vector estimation of a Gaussian mean from an initial model to an adapted model. The transfer vector is decomposed into a direction vector and a scaling factor. By using coarse class estimation for the direction vector, and by using fine class estimation for the scaling factor, we can represent accurate transfer vectors with a small number of parameters. Simple training algorithms for transfer vector estimation are analytically derived using the variational Bayes (VB) [7], maximum a posteriori (MAP) and ML methods. In addition, by using a VB solution, we control the fine and coarse classes according to the amount of adaptation data based on VB posteriors for model complexity. We demonstrate the effectiveness of our proposal in speaker adaptation experiments.

2. Coarse/Fine Training of transfer vectors

In this section, we introduce coarse/fine training of transfer vectors for mean vector parameters in acoustic model Gaussians. When the unknown data (adaptation data) is obtained from a new environment, we have to reconstruct an acoustic model from the obtained data and the initial model to deal with the new environment. A straightforward approach for reconstruction is to train acoustic model parameters based on ML by adding the obtained data to the original data. The new retrained parameter $\boldsymbol{\mu}_k^{new}$ for a mean vector in a Gaussian k is represented as follows:

$$\boldsymbol{\mu}_k^{new} = \frac{T_k^{ini}\boldsymbol{\mu}_k^{ini} + \zeta_k \boldsymbol{m}_k}{T_k^{ini} + \zeta_k}, \tag{1}$$

where T_k^{ini} and $\boldsymbol{\mu}_k^{ini}$ denote the occupation count and mean vector parameter of the initial data in Gaussian k, and ζ_k and $\boldsymbol{m}_k$ denote the occupation count and mean vector parameter of the new environment data. $\boldsymbol{\mu}_k^{new}$ corresponds to the interpolation vector between $\boldsymbol{\mu}_k^{ini}$ and $\boldsymbol{m}_k$, as shown in Figure 1. Here, we focus on the transfer vector $\boldsymbol{\mu}_k^{new} - \boldsymbol{\mu}_k^{ini}$ in Figure 1 and decompose it into a direction vector $\boldsymbol{\delta}_k$ and a scaling

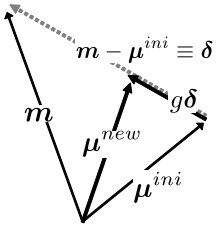

Figure 1: Definitions of a direction vector $\boldsymbol{\delta}$ and a scaling factor g.

factor g_k as follows:

$$\boldsymbol{\mu}_k^{new} = \boldsymbol{\mu}_k^{ini} + g_k \boldsymbol{\delta}_k : \begin{cases} \boldsymbol{\delta}_k & \equiv & \boldsymbol{m}_k - \boldsymbol{\mu}_k^{ini} \\ g_k & \equiv & \frac{\zeta_k}{T_k^{ini} + \zeta_k} \end{cases}, \quad (2)$$

The key point of our proposal is that we estimate $\boldsymbol{\delta}_{i_k}$ and g_{j_k} using different classes i_k and j_k from Gaussian class k, as follows:

$$\begin{aligned} \boldsymbol{\mu}_k^{new} &= \boldsymbol{\mu}_k^{ini} + \underline{g_k \boldsymbol{\delta}_k} \\ \longrightarrow \boldsymbol{\mu}_k^{new} &= \boldsymbol{\mu}_k^{ini} + \underline{g_{j_k} \boldsymbol{\delta}_{i_k}}, \end{aligned} \quad (3)$$

where classes i_k and j_k are different sets of Gaussians that both include Gaussian k. That is to say, *the direction vector and the scaling factor are to be tied across distinct sets of Gaussians*. The number of parameters for the scaling factor g is only one, and is much smaller than that for the direction vector $\boldsymbol{\delta}$, which equals the number of feature dimensions. This means that the estimation of the scaling factor g requires a much smaller amount of data than the estimation of the direction vector $\boldsymbol{\delta}$. Therefore, we can estimate the transfer vector for Gaussians even with a small amount of adaptation data by (i) estimating the direction vector from the large fraction of adaptation data assigned to tied Gaussians (coarse class estimation), and (ii) estimating the scaling factor from the small fraction of adaptation data assigned to an individual Gaussian (fine class estimation), as shown in Figure 2. We refer to this process as Coarse/Fine Training of transfer vectors (CFT).

Although the approach of CFT is similar to VFS (Vector Field Smoothing) [1], CFT's advantage is that the estimation is solved analytically by using the ML, MAP and VB approaches. The resulting training algorithms are as simple and robust as conventional Expectation Maximization (EM) methods. Especially in the VB approach, the tying class is automatically determined from adaptation data by considering the posterior distribution for model complexity. We introduce the analytical solutions for CFT in the next section.

3. Analytic solutions

Let $\boldsymbol{O} = \{\boldsymbol{o}^t \in \mathcal{R}^D : t = 1, ..., T\}$ be a set of D dimensional feature vectors. The complete data likelihood for a model

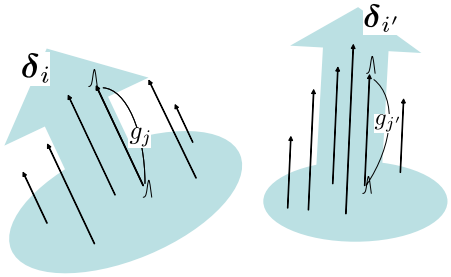

Figure 2: A mean transfer vector estimated by $\boldsymbol{\delta}_i$ and g_j.

parameter set Θ is then expressed as follows:

$$p(\boldsymbol{O}, S, V|\Theta) = \prod_t a^c_{s^{t-1}s^t} w^c_{s^t,v^t} \mathcal{N}(\boldsymbol{o}^t|\boldsymbol{\mu}^c_{s^t,v^t}, \Sigma^c_{s^t,v^t}), \quad (4)$$

where S is a set of sequences of HMM states, V is a set of sequences of Gaussian mixture components, and s^t and v^t denote the state and mixture components at frame t. Here, S and V are sets of discrete hidden variables. The parameter a denotes the state transition probability, and w is the weight factor of the Gaussian mixture. In addition, $\mathcal{N}(\cdot)$ denotes a Gaussian with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. c denotes the phoneme category index.

Here, we introduce the expectation form of complete data likelihood $p(\boldsymbol{O}, S, V|\Theta)$ for a posterior distribution for latent variables $p(S, V|\boldsymbol{O})$ as follows:

$$\langle \log p(\boldsymbol{O}, S, V|\Theta) \rangle_{p(S,V|\boldsymbol{O})} \propto \sum_{k,t} \zeta_k^t \log \mathcal{N}(\boldsymbol{o}^t|\boldsymbol{\mu}_k, \Sigma_k), \quad (5)$$

where k denotes all Gaussian indices over all HMM states in all phoneme categories, and $\langle f(y) \rangle_{p(y)}$ represents the expectation of $f(y)$ with respect to the distribution $p(y)$. This expectation form is used to calculate VB posteriors and MAP and ML parameters.

3.1. VB solution (CFT-VB)

VB is a powerful algorithm for practical posterior computation [7] and has been successfully applied to speech recognition using a method known as VBEC (Variational Bayesian Estimation and Clustering for speech recognition) [8, 9]. Here, we provide a VB solution for CFT. VB posteriors for model parameters $\widetilde{q}(\boldsymbol{\delta}_{i_k}|\boldsymbol{O})$ and $\widetilde{q}(g_{j_k}|\boldsymbol{O})$ are obtained as follows:

$$\begin{cases} \widetilde{q}(\boldsymbol{\delta}_{i_k}|\boldsymbol{O}) \propto p(\boldsymbol{\delta}_{i_k}) \exp \langle \log p(\boldsymbol{O}, S, V|\Theta) \rangle_{\widetilde{q}(S,V|\boldsymbol{O})\widetilde{q}(g_{j_k}|\boldsymbol{O})} \\ \widetilde{q}(g_{j_k}|\boldsymbol{O}) \propto p(g_{j_k}) \exp \langle \log p(\boldsymbol{O}, S, V|\Theta) \rangle_{\widetilde{q}(S,V|\boldsymbol{O})\widetilde{q}(\boldsymbol{\delta}_{i_k}|\boldsymbol{O})} \end{cases}, \quad (6)$$

where $p(\boldsymbol{\delta}_{i_k})$ and $p(g_{j_k})$ denote prior distributions for $\boldsymbol{\delta}_{i_k}$ and g_{j_k}, which are represented by Gaussians, and a tilde ($\widetilde{}$) is added to indicate variationally optimized values of functions. We assume that prior and posterior distributions for $\boldsymbol{\delta}_{i_k}$ and g_{j_k} are statistically independent of each other, as follows:

$$\begin{cases} p(\boldsymbol{\delta}_{i_k}, g_{j_k}) &= p(\boldsymbol{\delta}_{i_k}) p(g_{j_k}) \\ \widetilde{q}(\boldsymbol{\delta}_{i_k}, g_{j_k}|\boldsymbol{O}) &= \widetilde{q}(\boldsymbol{\delta}_{i_k}|\boldsymbol{O}) \widetilde{q}(g_{j_k}|\boldsymbol{O}) \end{cases}. \quad (7)$$

Then, by substituting Eqs. (3) and (5) into Eq. (6), we obtain a VB posterior for $\boldsymbol{\delta}_{i_k}$ as follows:

$$\widetilde{q}(\boldsymbol{\delta}_{i_k}|\boldsymbol{O}) = \mathcal{N}(\boldsymbol{\delta}_{i_k}|\widetilde{\boldsymbol{\alpha}}_{i_k}, \widetilde{\Omega}_{i_k}), \tag{8}$$

where $\widetilde{\boldsymbol{\alpha}}_{i_k}$ and $\widetilde{\Omega}_{i_k}$ are hyper-parameters for VB posteriors defined as:

$$\begin{aligned}\widetilde{\boldsymbol{\alpha}}_{i_k} &\equiv \widetilde{\Omega}_{i_k}\big((\Omega^0_{i_k})^{-1}\boldsymbol{\alpha}^0_{i_k} + \sum_{k\in i_k}\zeta_k\widetilde{u}_{j_k}(\Sigma_k)^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}^{ini}_k)\big)\\ \widetilde{\Omega}_{i_k} &\equiv \big((\Omega^0_{i_k})^{-1} + \sum_{k\in i_k}\zeta_k(\widetilde{v}_{j_k} + (\widetilde{u}_{j_k})^2)(\Sigma_k)^{-1}\big)^{-1}.\end{aligned} \tag{9}$$

Here, $\boldsymbol{\alpha}^0_{i_k}$ and $\Omega^0_{i_k}$ are hyper-parameters for a prior $p(\boldsymbol{\delta}_{i_k})$, and $\widehat{\boldsymbol{\mu}}_k \equiv \sum_t \zeta^t_k \boldsymbol{o}^t/\zeta_k$ is a mean vector for the adaptation data in Gaussian class k.

Similar to $\boldsymbol{\delta}_{i_k}$, VB posterior for g_{j_k} is also obtained by substituting Eqs. (3) and (5) into Eq. (6) as follows:

$$\widetilde{q}(g_{j_k}|\boldsymbol{O}) = \mathcal{N}(g_{j_k}|\widetilde{u}_{j_k}, \widetilde{v}_{j_k}), \tag{10}$$

where, $\widetilde{u}_{j_k}$ and $\widetilde{v}_{j_k}$ are hyper-parameters for VB posteriors defined as:

$$\begin{aligned}\widetilde{u}_{j_k} &\equiv \big((v^0_{j_k})^{-1}u^0_{j_k} + \sum_{k\in j}\zeta_k\widetilde{\boldsymbol{\alpha}}'_{i_k}(\Sigma_k)^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}^{ini}_k)\big)\widetilde{v}_{j_k}\\ \widetilde{v}_{j_k} &\equiv \big((v^0_{j_k})^{-1} + \sum_{k\in j}\zeta_k \mathrm{tr}((\widetilde{\boldsymbol{\alpha}}_{i_k}\widetilde{\boldsymbol{\alpha}}'_{i_k} + \widetilde{\Omega}_{i_k})(\Sigma_k)^{-1}\big)^{-1}.\end{aligned} \tag{11}$$

Here, $u^0_{i_k}$ and $v^0_{i_k}$ are hyper-parameters for a prior $p(g_{i_k})$, and $'$ and tr denote the transpose and the trace of the matrix.

The VB objective function $\mathcal{F}$, which is proportional to the posterior probability for a model complexity, can be calculated by

$$\mathcal{F} = \left\langle \log \frac{p(\boldsymbol{O}, S, V|g, \boldsymbol{\delta})p(g, \boldsymbol{\delta}|\boldsymbol{O})}{\widetilde{q}(S, V|\boldsymbol{O})\widetilde{q}(g, \boldsymbol{\delta}|\boldsymbol{O})} \right\rangle_{\widetilde{q}(S,V|\boldsymbol{O}),\widetilde{q}(g,\boldsymbol{\delta}|\boldsymbol{O})}, \tag{12}$$

using $\widetilde{q}(\boldsymbol{\delta}_{i_k}|\boldsymbol{O})$ and $\widetilde{q}(g_{j_k}|\boldsymbol{O})$. An appropriate model structure is selected by maximizing the VB objective function $\mathcal{F}$ with respect to a model structure [7–9].

3.2. MAP solution (CFT-MAP)

Next, we introduce the MAP solution for CFT. MAP estimates of $\boldsymbol{\delta}_{i_k}$ and g_{j_k} are obtained by constructing an auxiliary function (known as Q function) from Eq. (5) and priors, and by differentiating it with respect to $\boldsymbol{\delta}_{i_k}$ and g_{j_k} [2]. Analytic solutions are as follows:

$$\begin{aligned}\boldsymbol{\delta}^{MAP}_{i_k} =& \big((\Omega^0_{i_k})^{-1} + \sum_{k\in i_k}\zeta_k(g^{MAP}_{j_k})^2(\Sigma_k)^{-1}\big)^{-1}\\ &\cdot\big((\Omega^0_{i_k})^{-1}\boldsymbol{\alpha}^0_{i_k} + \sum_{k\in i_k}\zeta_k g^{MAP}_{j_k}(\Sigma_k)^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}^{ini}_k)\big)\\ g^{MAP}_{j_k} =& \frac{(v^0_{j_k})^{-1}u^0_{j_k} + \sum_{k\in j}\zeta_k(\boldsymbol{\delta}^{MAP}_{i_k})'(\Sigma_k)^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}^{ini}_k)}{(v^0_{j_k})^{-1} + \sum_{k\in j}\zeta_k(\boldsymbol{\delta}^{MAP}_{i_k})'(\Sigma_k)^{-1}\boldsymbol{\delta}^{MAP}_{i_k}}.\end{aligned} \tag{13}$$

We focus on the obtained scaling factor $g^{MAP}_{j_k}$ and the mean of the VB posterior for a scaling factor $\widetilde{u}_{j_k}$ and discuss the analytic solutions. When ζ_k becomes small, $g^{MAP}_{j_k}$ and $\widetilde{u}_{j_k}$ approach $u^0_{j_k}$. Therefore, by setting $u^0_{j_k}$ to a small value, $g^{MAP}_{j_k}$ and $\widetilde{u}_{j_k}$ approach that small value, and their transfer vectors $g\boldsymbol{\delta}$ also decrease. Conversely, when ζ_k becomes large, $g^{MAP}_{j_k}$ and $\widetilde{u}_{j_k}$ approach 1. These limits for large and small amounts of data show the validity of the solutions; i.e., $\boldsymbol{\nu}^{new}$ does not move far from $\boldsymbol{\nu}^{ini}$ when the amount of data is small, and $\boldsymbol{\nu}^{new}$ approaches $\boldsymbol{m}$ when the amount of data is large.

3.3. ML solution (CFT-ML)

Finally, we introduce an ML solution for CFT, as follows:

$$\begin{aligned}\boldsymbol{\delta}^{ML}_{i_k} =& \big(\sum_{k\in i}\zeta_k(g^{ML}_{j_k})^2(\Sigma_k)^{-1}\big)^{-1}\\ &\cdot\sum_{k\in i}\zeta_k g^{ML}_{j_k}(\Sigma_k)^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}^{ini}_k)\\ g^{ML}_{j_k} =& \frac{\sum_{k\in j}\zeta_k(\boldsymbol{\delta}^{ML}_{i_k})'(\Sigma_k)^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}^{ini}_k)}{\sum_{k\in j}\zeta_k(\boldsymbol{\delta}^{ML}_{i_k})'(\Sigma_k)^{-1}\boldsymbol{\delta}^{ML}_{i_k}}.\end{aligned} \tag{14}$$

Although an ML solution is the simplest, it has a singular point at $\sum_{i\in k}\zeta^k = 0$, that causes incorrect estimation.

4. Experiments

We conducted experiments to evaluate the effectiveness of the proposed CFT adaptation. The experiments examined how CFT works with varying amounts of adaptation data in comparison to conventional MAP adaptation using initial models as priors [2]. Among the three derivations of CFT presented here, we chose CFT-VB, which is a complete version of CFT that includes model selection capability based on a VB objective function $\mathcal{F}$. The tied-Gaussian classes for the vector $\boldsymbol{\delta}_{i_k}$ are organized so that each class corresponds to a clustered state in a triphone HMM, and the scaling factor g_{j_k} is estimated for the individual Gaussian class, i.e., $g_{j_k} \rightarrow g_k$. The variance parameters are kept constant during both CFT-VB and MAP estimations. We performed the experiments under the conditions shown in Tables 1 and 2.

The total training data for initial models consisted of about 4,800 Japanese words spoken by 50 males. The total adaptation and recognition data consisted of 1,300 Japanese words spoken by one male who was not included in the initial model training. We divided a set of isolated word data into adaptation and recognition data. The total adaptation data consisted of 1,000 words, while the remaining 300 words were assigned to the recognition data. Several subsets were randomly extracted from the adaptation data set, and each of these subsets was used to construct a set of adapted acoustic models. As a result, about 20 sets of adapted acoustic models for several amounts of adaptation data were prepared. In the initial model training, we constructed 324 speaker-independent triphone HMM states, clustered using a pho-

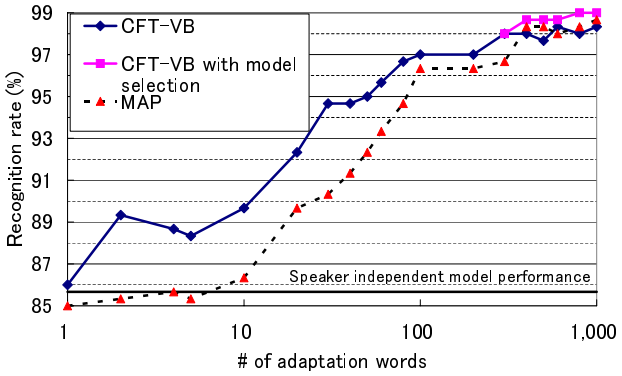

Figure 3: Recognition rate for supervised speaker adaptation task with varying amounts of adaptation data.

netic decision tree method. The output distribution in each state is represented by an 8-component Gaussian mixture model.

Figure 3 compares the recognition results obtained by the CFT-VB and MAP methods for several amounts of adaptation data with the performance for a non-adapted speaker independent model (85.7 %). For a small amount of data (fewer than 100 words), CFT-VB was superior to MAP by up to 4 %. For a large amount of data (more than 100 words), CFT-VB was comparable to MAP. These results using large and small amounts of data support the effectiveness of CFT, which employs both coarse and fine estimation to improve adapted models for any amount of data.

Next, we utilized VB model selection in the adaptation experiments. The estimation class of the scaling factor was kept in the Gaussian class. For the estimation of the direction vector $\boldsymbol{\delta}_{i_k}$, we provided the previously described clustered state (coarse) class and the Gaussian (fine) class as candidates for model selection. The appropriate class for $\boldsymbol{\delta}_{i_k}$ was selected from the coarse and fine classes for every amount of data by using a VB objective function $\mathcal{F}$. The class of the direction vector $\boldsymbol{\delta}_{i_k}$ was automatically changed from the Gaussian class to the clustered state class at 400 words adaptation using CFT-VB with model selection, as shown in Figure 3. This result means that the amount of data per parameter became sufficiently large at 400 words, so the class was automatically changed from coarse to fine. The performance obtained from fine class estimation was superior to MAP, even for a large amount of training data, as shown in Figure 3. The improvement comes from the fine model structure selected by CFT-VB, which provided a more exact model representation than MAP using the additional scaling factor parameters. Thus, CFT-VB achieved better performance than MAP with a large amount of data due to the accurate VB model selection.

As a result of coarse/fine training, and the use of VB model selection, CFT was superior to MAP for any amount of adaptation data.

Table 1: Acoustic Conditions

Sampling rate	16 kHz (quantization 16 bit)
Feature vector	12 - order MFCC with Δ MFCC
Window	Hamming
Frame size/shift	25/10 ms

Table 2: Acoustic model structure

# of states	3 (left to right)
# of phoneme categories	27
# of clustered states	324
Output distribution	8 components GMM

5. Summary

In this paper we proposed a novel acoustic model adaptation technique based on Coarse/Fine Training of transfer vectors (CFT) and applied CFT to a supervised speaker adaptation task. CFT was superior to the conventional MAP adaptation due to the use of coarse class estimation for small amounts of training data. In addition, by utilizing the VB model selection, CFT was superior to MAP adaptation even for a large amount of training data due to accurate model selection. CFT is a simple and powerful adaptation technique, which we will apply to unsupervised, on-line, and incremental adaptation tasks in the future.

6. References

[1] K. Ohkura, M. Sugiyama, and S. Sagayama, "Speaker adaptation based on transfer vector field smoothing with continuous mixture density HMMs," in *Proc. ICSLP1992*, 1992, vol. 1, pp. 369–372.

[2] J-L. Gauvain and C-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on SAP*, vol. 2, pp. 291–298, 1994.

[3] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[4] M. Tonomura, T. Kosaka, and S. Matsunaga, "Speaker adaptation based on transfer vector field smoothing using maximum a posteriori probability estimation," in *Proc. ICASSP1995*, 1995, vol. 1, pp. 688–691.

[5] J. Takahashi and S. Sagayama, "Vector-field-smoothed Bayesian learning for incremental speaker adaptation," in *Proc. ICASSP1995*, 1995, vol. 1, pp. 696–699.

[6] K. Shinoda and C-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. on SAP*, vol. 9, pp. 276–287, 2001.

[7] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," in *Proc. UAI 15*, 1999.

[8] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, *Application of variational Bayesian approach to speech recognition*, NIPS 2002, MIT Press, 2003.

[9] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Variational Bayesian estimation and clustering for speech recognition," *IEEE Trans. on SAP*, 2004, (to appear).