6-2000

# On-Line Bayesian Speaker Adaptation By Using Tree-Structured Transformation and Robust Priors

Shaojun Wang
*Wright State University - Main Campus*, shaojun.wang@wright.edu

Yunxin Zhao

## Repository Citation

Wang, S., & Zhao, Y. (2000). On-Line Bayesian Speaker Adaptation By Using Tree-Structured Transformation and Robust Priors. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2*, 977-980.
https://corescholar.libraries.wright.edu/knoesis/1022

# ON-LINE BAYESIAN SPEAKER ADAPTATION USING TREE-STRUCTURED TRANSFORMATION AND ROBUST PRIORS

*Shaojun Wang*[1] *Yunxin Zhao*[2]

Beckman Institute and Dept. of ECE, University of Illinois, Urbana, IL 61801[1]
Dept. of CECS, University of Missouri, Columbia, MO 65211[2]
*swang@ifp.uiuc.edu* *zhao@cecs.missouri.edu*

## ABSTRACT

This paper presents new results by using our recently proposed on-line Bayesian learning approach for affine transformation parameter estimation in speaker adaptation. The on-line Bayesian learning technique allows updating parameter estimates after each utterance and it can accommodate flexible forms of transformation functions as well as prior probability density function. We show through experimental results the robustness of heavy tailed priors to mismatch in prior density estimation. We also show that by properly choosing the transformation matrices and depths of hierarchical trees, recognition performance improved significantly.

## 1. INTRODUCTION

On-line speaker adaptation is a practical method of coping with talker variabilities to improve accuracy of speech recognition. Several approaches appeared in the literature [3, 6, 10, 17, 18] for on-line speaker adaptation. One approach [17, 18] used EM algorithm or segmental $k$-means training algorithm sequentially to on-line test speech to accomplish unsupervised learning of model parameters. Since the accumulated sufficient statistics from each past utterance are computed using the model parameters updated at that time, the model parameter estimates are not as accurate as batch training.

Another approach [8] uses an *incremental* version of the EM algorithm proposed in [14]. In incremental EM approach, when the complete-data likelihood has the regular exponential-family form, the E step reduces to incrementally computing and maintaining the conditional sufficient statistics, and M step to computing the ML/MAP estimates given these conditional sufficient statistics. In the $k$-th iteration, the conditional sufficient statistics of current observation are computed using the $k$-1th model estimates, others are unchanged, as opposed to the batch version, and after each M step the conditional sufficient statistics of all the training data are recalculated using the lastest parameter estimates. In incremental EM training, eventhough likelihood is not guaranteed to increase as in the batch EM algorithm [2], its convergence is recently proved by Csiszar's alternating minimization procedure [9]. However, as pointed by Digalakis [6], this incremental EM algorithm is not an on-line algorithm since multiple passes through the data are performed and storage of current value of the conditional sufficient statistics from each past observations is required. Digalakis [6] suggested modifications to the incremental EM algorithm to enable its on-line computation. However, the convergence of this modified incremental EM is still open.

On-line quasi-Bayes learning [10] approximates the successive posterior distributions by the "closest" tractable distribution within a given class $\mathcal{P}$, under the criterion that both distributions have the same mode. Then the EM algorithm is applied and the hyperparameters of the approximate posterior distribution and model parameters are incrementally updated. Empirical evidence showed that the quasi-Bayes algorithm does converge to a reasonable solution in terms of improving recognition rate and it has a similar behavior with the batch MAP algorithm [10].

In [16], we proposed an on-line Bayesian transformation algorithm that uses a hierarchical tree structure to control the degree of transformation tying. The underlying theoretical framework is the recursive Bayesian learning we developed recently. This technique allows updating parameter estimates after each utterance and it can accommodate flexible forms of transformation functions as well as prior probability density functions. Through speaker adaptation, recognition accuracy was consistently improved for increasing number of adaptation data, and the performance was shown to be superior to existing on-line adaptation methods. In the current work, we further investigate the choices on the form of prior density function, transformation matrices, depth of hierarchical trees, and demonstrate their significant impact on recognition performance. We show that heavy tailed priors are more robust for speaker adaptation, and the form of transformation matrix and the tree size should be adaptive to data size.

## 2. ON-LINE BAYESIAN LEARNING FOR TREE-STRUCTURED TRANSFORMATION PARAMETERS

In this section, we briefly summarize recursive Bayesian learning for tree-structured transformation parameters [16]. Assume that $\underline{o}^k = \{\underline{o}_1, \cdots, \underline{o}_k\}$ are $k$ successively independent blocks of speech feature vectors and are described by a probability density function with parameter $\eta$. Let $Q_{\underline{o}^{k+1}}(\eta, \eta^{(k)})$ denote the auxiliary function of log likelihood as defined in EM algorithm [4, 13] and $p(\eta|\phi)$ be the prior pdf of $\eta$ with parameter $\phi$. Define $R_{\underline{o}^{k+1}}(\eta, \eta^{(k)}) = Q_{\underline{o}^{k+1}}(\eta, \eta^{(k)}) + logp(\eta|\phi)$, which becomes the auxiliary function of log posterior likelihood. Let $\ell_k(\eta, \eta^{(k)}) = R_{\underline{o}^{k+1}}(\eta, \eta^{(k)}) - R_{\underline{o}^k}(\eta, \eta^{(k)}) = Q_{\underline{o}_{k+1}}(\eta, \eta^{(k)})$. Then a recursive estimation algorithm can be derived as

$$\eta^{(k+1)} = \eta^{(k)} + \varepsilon_k H(\underline{o}^{k+1}, \eta^{(k)})^{-1} \tfrac{1}{k+1} \tfrac{\partial \ell_k(\eta, \eta^{(k)})}{\partial \eta}]|_{\eta=\eta^{(k)}} \qquad (1)$$

where recursive computation of $H(\underline{o}^{k+1}, \eta^{(k)})$ is approximated as $\frac{1}{k+1}[\sum_{i=1}^{k+1} I_C(\underline{o}_i|\eta^{(i)}) + I_p(\eta^{(k)})]$, $I_C(\underline{o}_i|\eta^{(i)})$ is the conditional expectation of the complete-data information matrix given $\underline{o}_i$ [13], $I_p(\eta)$ is prior information matrix, i.e., negative Hessian matrix of log $p(\eta|\phi)$. The optimal choice of $\varepsilon_k$ at each step is determined by line search [12] along the direction of $H(\underline{o}^{k+1}, \eta^{(k)})^{-1} \frac{1}{k+1} \frac{\partial \ell_k(\eta, \eta^{(k)})}{\partial \eta}]|_{\eta=\eta^{(k)}}$ to maximize $\frac{1}{k+1} R_{\underline{o}^{k+1}}(\eta, \eta^{(k)})$, which is approximated by $\frac{1}{k+1} \sum_{i=1}^{k+1} [Q_{\underline{o}_i}(\eta, \eta^{(i-1)}) + I_p(\eta^{(k)})]$.

Continuous density hidden Markov models of isolated words are adapted with state-dependent observation pdfs being Gaussian mixture densities. Affine transformation [5] is applied to the observations in the feature space, i.e., $\underline{O}_t = A\underline{Y}_t + b$, and it is equivalent to a constrained model space transformation on both the mean vectors and covariance matrices of Gaussian densities. For the $i$th state and

0-7803-6293-4/00/$10.00 ©2000 IEEE.

$m$th mixture, the transformation has the form of

$$\hat{\lambda} = G_{\eta^{(k)}} = \left(c_{i,m}, A_c^{(k)}\mu_{i,m} + b_c^{(k)}, A_c^{(k)}\Sigma_{i,m}(A_c^{(k)})^T\right) \quad (2)$$

The transformation function $G_\eta(\cdot)$ has $C$ clusters with $\eta^{(k)} = \{(A_c^{(k)})^{-1}, b_c^{(k)}\}, c = 1, \cdots, C$, and each $\lambda_{i,m} = (\mu_{i,m}, \Sigma_{i,m})$ is assumed to be labeled by a cluster membership $\Omega_c$. Details of derivations can be found in [16].

A hierarchical tree of the entire set of HMM Gaussian parameters is employed to dynamically control the transformation tying and the prior knowledge of Gaussian parameters is incorporated to estimate the transformation parameters. The Gaussian densities are clustered by using the binary split $K$-means algorithm with a divergence measure [3, 16]. In [3], a *bottom-up* strategy is proposed to automatically search for the transformation parameters of each Gaussian mixture component, the computational cost is $O(G\log d)$, where $G$ is the number of Gaussian components and $d$ the the depth of the hierarchical tree. In [16], we use a more efficient strategy, i.e., white-black tree based *bottom-up top-down* strategy, for on-line Bayesian learning of affine transformation parameters, where a *bottom-up* procedure is first used to perform on-line Bayesian learning of transformation parameters $\eta_c^{(k)}$ for the nodes containing adaptation data, and a *top-down* procedure is next used to perform transformations on all the HMM's Gaussian mixture components, and the computation cost is $O(G)$.

## 3. SELECTION AND ESTIMATION OF PRIOR PROBABILITY DENSITY FUNCTION

Within a Bayesian framework, prior distributions can be assigned to the transformation parameters $\eta$. The choices of prior distribution are often quite arbitrary and depends on the type of prior knowledge available. Different choices of the priors in principle will lead to different robustness properties of the learning procedure. Because there is no particular reason to believe that a postulated prior is true, one would like the estimator to be relatively insensitive to departures of the prior from the assumed form. This is a problem of Bayesian robustness with respect to the prior [1]. Standard choices such as the exponential family and conjugate prior, are known to be nonrobust in various ways: models in the exponential family are very sensitive to outliers in the data, and conjugate priors can have a pronounced effect on the answers even if the data is in conflict with the specified prior information [1]. On the other hand, the distributions with heavy tails, such as $\alpha$-stable, generalized Gaussian and Student-$t$ families, tend to be much more robust than the standard choices.

In the current work, we adopt the generalized Gaussian density priors (GGD), which have the form

$$p(\eta|\phi) = \frac{\gamma\Gamma^{\frac{1}{2}}(3/\gamma)}{2\Gamma^{\frac{3}{2}}(1/\gamma)}|\Sigma_\eta|^{-1/2}exp[-(\rho(\gamma)f(\eta|\phi))^{\gamma/2}] \quad (3)$$

where $f(\eta|\phi) = (\eta - \mu_\eta)^T\Sigma_\eta^{-1}(\eta - \mu_\eta)$, $\rho(\gamma) = \frac{\Gamma(3/\gamma)}{\Gamma(1/\gamma)}$ with $\gamma$ the shape parameter, and $\Gamma(\cdot)$ the Gamma function. The pdf is also known as the power exponential distribution, or $\alpha$-Gaussian etc. The GGD model contains the Gaussian and Laplacian density functions as special cases when using $\gamma = 2$ and $\gamma = 1$, respectively. For decreasing values of $\gamma$, the tails of the distribution become increasingly flat. The pdf exhibits an algebraic singularity at $\eta = \mu_\eta$ for $0 < \gamma < 1$, and as $\gamma$ goes to zero, $p(\mu_\eta)$ goes to infinity. Fig. 1 illustrates the distributions corresponding to $\gamma = 0.7, 1, 2$ and 5.

The prior information matrix is given as

$$\begin{aligned} I_p(\eta) = -\partial^2 logp(\eta|\phi)/\partial\eta\partial\eta^T = \rho(\gamma)^{\gamma/2}[\gamma(f(\eta|\phi)^{\frac{\gamma}{2}-1}\Sigma_\eta^{-1} \\ +2\gamma(\tfrac{\gamma}{2} - 1)f(\eta|\phi)^{\frac{\gamma}{2}-2}(\Sigma_\eta^{-1}(\eta - \mu_\eta))(\Sigma_\eta^{-1}(\eta - \mu_\eta))^T] \end{aligned} \quad (4)$$

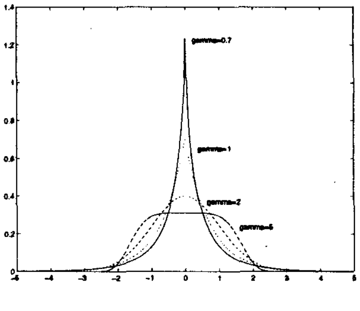

**Figure 1.** *GGD's with unit variance and shape parameters* $\gamma = 0.7, 1, 2,$ *and* 5

Empirical Bayes approach has been commonly used to estimate the parameters $\phi$ of prior densities [1, 7]. In the empirical Bayes approach, speaker independent training data set $O$ for estimating hyperparameters $\phi_c$ can be divided into different subsets $O_1, \cdots, O_S$ that correspond to $S$ different speakers and each has the transformation parameter $\eta_{c,s}$. The parameter $\eta_{c,s}$ is accounted as random observations generated by a common prior distribution $p(\eta_c|\phi_c)$. The marginal distribution of training data $O$ can be written as

$$p(O|\phi_c) = \int \prod_{s=1}^S p(O_s|\eta_{c,s})p(\eta_{c,s}|\phi_c)d\eta_{c,s} \quad (5)$$

There are two methods for parameter estimation of prior density, namely moment approach and type-II ML approach [1]. The moment approach applies when it is possible to relate prior moments to moments of the marginal distribution $p(O|\phi_c)$. This method is often difficult to apply except in few simple cases as provided in [3, 10]. In type-II maximum likelihood approach, a maximum likelihood estimate, $\hat{\phi}_c$, is obtained by maximizing $p(O|\phi_c)$, and $p(\eta_c|\hat{\phi}_c)$ is called type-II maximum likelihood prior [1]. However, $\hat{\phi}_c$ is rather difficult to obtain due to the integration in Eqn. (5). Although EM algorithm can be applied as suggested in [4], it is often difficult to derive an explicit expression of solution in the E step of EM.

To alleviate the problem, we attempt to maximize an approximation of the integral in Eqn. (5). We assume that for any $\phi_c$, $p(O, \eta_{c,s}|\phi_c)$ is sharply peaked at $\hat{\eta}_{c,s} = argmax_{(\eta_{c,s})}p(O, \eta_{c,s}|\phi_c)$ and we try to find $\hat{\phi}_c$ to maximize $p(O, \hat{\eta}_{c,s}|\phi_c)$. This leads to an alternative maximization procedure over $\eta_c$ and $\phi_c$ as suggested in [7], i.e.

$$\eta_{c,s}^{(k)} = argmax_{(\eta_{c,s})}p(O_s, \eta_{c,s}|\phi_c^{(k)}), \; s = 1, \cdots, S \quad (6)$$

$$\begin{aligned} \phi_c^{(k+1)} &= argmax_{(\phi_c)}p(O, \hat{\eta}_{c,s}|\phi_c) \\ &= argmax_{(\phi_c)}\prod_{s=1}^S p(\eta_{c,s}^{(k)}|\phi_c) \end{aligned} \quad (7)$$

Here Eqn. (6) is used in the posterior estimate of $\eta_{c,s}$ and it can be solved by the batch EM gradient method in [16]. For Eqn. (7), there is no closed form solution, except for the case $\gamma = 2$ (Gaussian). For each given $\eta_{c,s}^{(k)}$, $\phi_c^{(k+1)}$ can be iteratively solved as

$$\mu_\eta^{(k+1)}(i+1) = \frac{\sum_{s=1}^S f(\eta_s^k|\phi^{(k+1)}(i))^{\gamma/2-1}\eta_s^k}{\sum_{s=1}^S f(\eta_s^k|\phi^{(k+1)}(i))^{\gamma/2-1}} \quad (8)$$

$$\Sigma_\eta^{(k+1)}(i+1) = \frac{\sum_{s=1}^S f(\eta_s^k|\phi^{(k+1)}(i))^{\gamma/2-1}(\eta_s^k - \mu_\eta^{(k+1)}(i))(\eta_s^k - \mu_\eta^{(k+1)}(i))^T}{\sum_{s=1}^S f(\eta_s^k|\phi^{(k+1)}(i))^{\gamma/2-1}} \quad (9)$$

where index $i$ denotes the iteration number. Note that when $\gamma = 2$, $f(\eta_s|\phi)^{\gamma/2-1} = 1$ and the above equations are the explicit solution of maximum likelihood estimation.

## 4. EXPERIMENTAL RESULTS

The on-line Bayesian learning approach is applied to online speaker adaptation using a vocabulary of 26-letter English alphabet. Two severely mismatched speech databases , the OGI ISOLET and the TI46, were used for evaluating the adaptation algorithm. A full description of these two corpora can be found in [10]. For speaker independent training, the OGI ISOLET database was used. It consists of 150 speakers, each speaking the letter twice. For on-line Bayesian adaptation and testing, the English alphabet subset of the TI46 isolated word corpus was used. Among the 16 talkers, data from 4 males were incomplete. Therefore, only 12 speakers (4 males and 8 females) were used in this study. Each person uttered each of the letters 26 times, where ten were collected in the same session and the remaining 16 were collected in 8 different sessions with 2 in each session. For each person and each letter, we divided equally those 16 tokens collected in eight different sessions into two parts, one for adaptive training, another for testing. Each letter was modeled by a single left-to-right five-state CDHMM. Each state had a Gaussian mixture density of four components with each component density having a diagonal covariance matrix. The speech feature was extracted based on a tenth-order LPC analysis, where the feature components were 12 cepstral coefficients, a normalized log energy and their first time derivatives.

A number of experiments were conducted, which included a comparison among Generalized Gaussian prior models with different $\gamma$'s, a comparison among using diagonal, block diagonal and full transformation matrices, a comparison among the use of different depths of hierarchical tree. By default, we used six layered hierarchical tree and Gaussian prior in the experiments.

I. Robustness of Priors

We investigated the robustness of generalized Gaussian prior models. Here we used block diagonal matrix and the depth of hierartical tree was 6. The recognition performace by using three values of the shape parameter $\gamma = 2, 1, 0.7$ are shown in Fig. 2. As the results indicate, the recognition performance was improved when using priors with heavy tails than using standard Gaussian. There was little difference between the results for $\gamma = 1$ and $\gamma = 0.7$.

II. Full, Block Diagonal and Diagonal Transformation Matrices

A transformation matrix can be chosen as full, diagonal or block diagonal. The use of block diagonal matrix is based on the assumptation that a seperate transformation can be used for each type of speech features, including cepstral coefficients, energy, first-order time derivatives, resulting in a block diagonal transformation matrix in which parameter correlation is considered only within the same set of features. The choice of the transformation matrix structure is in general a trade-off between the number of parameters to be estimated and the amount of adaptation data required. The problem is refered to as *model parameterization complexity*. Previously, Leggetter and Woodland [11] investigated the effect on recognition performance by taking transformation matrix to be either full or diagonal in maximum likelihood linear regression batch adaptation approach. Here we investigate the effect in on-line Bayesian learning of tree-structured affine transformation parameters.

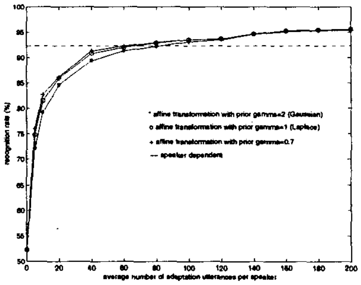

**Figure 2.** *Performance of recognition results by GGD Priors*

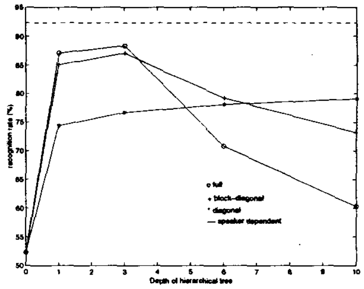

**Figure 3.** *Full, block-diagonal, and diagonal matrix using 10 adaptation utterences per speaker*

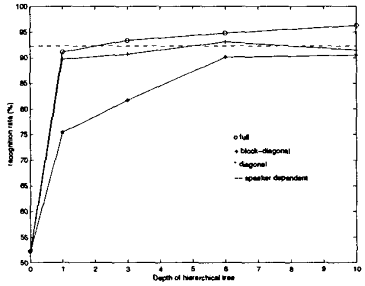

**Figure 4.** *Full, block-diagonal, and diagonal matrix using 100 adaptation utterences per speaker*

We tested the effect of model parameterization complexity on recognition performance by using full, block diagonal as well as diagonal matrices while adaptation was using different depths of hierarchical tree, where the total adaptation data were 10 utterances per speaker in one case and were 100 in another.

Fig. 3 gives the recognition performance when using the small amount of adaptation data. All transformations provided improvements over the initial model, but the effect of diagonal matrices was limited. The full matrices gave a substantial improvement when using one or three layered trees. However, as the depth of the tree was increased, the amount of data allocated to each leaf became small and the matrices were poorly estimated, and thus the performance of full matrices droped rapidly. With diagonal matrices, as deeper trees were used the performance gradually increased; however, this effect was very small and using 500 diagonal matrices was only 5.0% better than using one diagonal matrix. The amount of data needed to estimate a diagonal matrix is much smaller than that of a full matrix. This indicates that deeper tree can be used for diagonal matrices than for full matrices with the same amount of data. The results show that increasing the depth of the tree did improve performance, but the performance never reached that of the full matrix. It is clear that the off-diagonal terms accounting for the interdependencies between elements of feature vectors were important.

Fig. 4 gives the recognition performance when using the large amounts of adaptation data. In this case, the amount of data allocated to each leaf were abundant and the matrices were well estimated. Since full transformation matrices

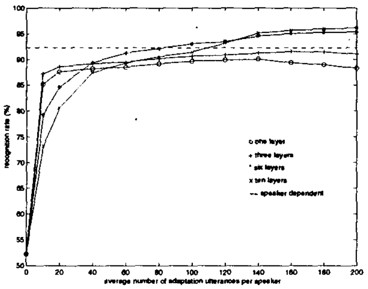

**Figure 5.** *Performance of recognition results by various depths*

take into accounts of interdependencies among elements of the feature vectors, it gave substantially better performance than those of diagonal transformation matrices.

III. Depth of Hierarchical Tree

We investigated the effects of hierarchical tree with different depths, which is termed as the problem of *model structure complexity*. Previously, Shinoda and Watanabe [15] investigated this problem for bias transformation by using batch approach. Here we investigate this problem in online Bayesian learning of tree-structured affine transformation parameters. Fig. 5 gives the recognition results of online learning of block-diagonal transformation matrices with various tree depths while varying the number of adaptation utterances. We observe that when the number of adaptation data was small, good recognition performances were achieved by trees with shallow depths, and trees with deeper depths gave poor performance, since deeper tree structure overfitted the limited training data. As more adaptation data was presented, the performance of all trees gradually improved, but the trees with shallow depths improved little, which indicates underfitting the adaptation data. As sufficient amount of data was used, good recognition performance was achieved by trees with deeper depths. From the above results, we see that a critical issue is obtaining right-sized trees, i.e., trees which neither underfit nor overfit the adaptation data.

## 5. DISCUSSION AND CONCLUSION

The proposed on-line Bayesian learning technique allows updating parameter estimates after each utterance and can accommodate flexible forms of transformation functions as well as prior probability density functions. In this work, we suggested using GGD as priors for on-line Bayesian learning of tree-structured transformation of Gaussian densities of hidden Markov models. It was found that heavy tailed prior density functions gave better recognition performance and thus are more robust to mismatch in prior estimation.

From the experimental results, we can see that in order to best use the adaptation data, information criteria of model selection needs to be used. The issue of how to accomodate various information criteria into on-line Bayesian adaptation to control both model structural complexity and parameterization complexity is a further research direction we are going to explore.

## 6. ACKNOWLEDGEMENT

This work was supported in part by the National Science Foundation under grant NSF IRI-95-02074.

## REFERENCES

[1] J. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer, 1985

[2] W. Byrne, A. Gunawardana and S. Khudanpur, "Comments on Efficient Training Algorithms for HMM's Using Incremental Estimation," submitted to *IEEE Trans. on Speech and Audio Processing*, 1999

[3] J. Chien, "On-Line Hierarchical Transformation of Hidden Markov Models for Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 7, No. 6, pp. 656-668, 1999

[4] A. Dempster, N. Laird and D. Rubin, "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm (with discussion)," *Journal of Royal Statistical Society, Series B*, Vol. 39, pp 1-38, 1977

[5] V. Digalakis, D. Rtischev and L. Neumeyer, "Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures," *IEEE Trans. on Speech and Audio Processing*, Vol. 3, pp. 357-366, September 1995

[6] V. Digalakis, "On-Line Adaptation of Hidden Markov Models Using Incremental Estimation Algorithms," *IEEE Trans. on Speech and Audio Processing*, pp. 253-261, May 1999

[7] J. Gauvain and C. Lee, "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 2, pp. 291-298, April 1994

[8] Y. Gotoh, M. Hochberg and H. Silverman, "Efficient Training Algorithms for HMM's Using Incremental Estimation," *IEEE Trans. on Speech and Audio Processing*, Vol. 6, No. 6, pp. 291-298, November 1998

[9] A. Gunawardana and W. Byrne, "Convergence of EM Variants," CLSP Research Note, No. 32, Johns Hopkins University, 1999

[10] Q. Huo and C. Lee, "On-Line Adaptive Learning of the Continuous Density Hidden Markov Model Based on Approximate Recursive Bayes Estimate," *IEEE Trans. on Speech and Audio Processing*, Vol. 5, No. 2, pp. 161-172, March 1997

[11] 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

[12] D. Luenberger, *Linear and Nonlinear Programming*, Addison-Wesley, 1984

[13] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, John Wiley & Sons, New York, 1997

[14] R. Neal and G. Hinton "A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants," in *Learning in Graphical Models* edited by M. Jordan, pp. 355-368, Kluwer Academic Publishers, 1998

[15] K. Shinoda and T. Watanabe, "Speaker Adaptation with Autonomous Model Complexity Control by MDL Principle," *IEEE Proceedings of International Conference on Acoustic, Speech and Signal Processing*, pp. 717-720, 1996

[16] S. Wang and Y. Zhao, "On-Line Bayesian Tree-Structured Transformation of Hidden Markov Models for Speaker Adaptation," *IEEE Workshop on Automatic Speech Recognition and Understanding*, December 12-15, Keystone, Colorado, 1999

[17] G. Zavaliagkos, R. Schwartz and J. Makhoul, "Batch, Incremental and Instantaneous Adaptation Techniques for Speech Recognition," *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 676-679, May 1995

[18] Y. Zhao, "Self-Learning Speaker and Channel Adaptation Based on Spectral Variation Source Decomposition," *Speech Communication*, Vol. 18, pp. 65-77, 1996