ON-LINE ADAPTIVE LEARNING OF THE CORRELATED CONTINUOUS DENSITY HIDDEN MARKOV MODELS FOR SPEECH RECOGNITION

Qiang Huo[†] and Chin-Hui Lee[‡]

[†]ATR Interpreting Telecommunications Research Labs., 2-2 Hikaridai, Seika-cho Soraku-gun, Kyoto 619-02, Japan
[‡]Multimedia Communication Research Lab, Bell Laboratories, Murray Hill, NJ 07974, USA

ABSTRACT

We extend our previously proposed quasi-Bayes adaptive learning framework to cope with the correlated continuous density hidden Markov models with Gaussian mixture state observation densities in which all mean vectors are assumed to be correlated and have a joint prior distribution. A successive approximation algorithm is proposed to implement the correlated mean vectors' updating. As an example, by applying the method to on-line speaker adaptation application, the algorithm is experimentally shown to be asymptotic convergent as well as being able to enhance the efficiency and the effectiveness of the Bayes learning by taking into account the correlation information between different models. The technique can be used to cope with the time-varying nature of some acoustic and environmental variabilities, including mismatches caused by changing speakers, channels, transducers, environments and so on.

1. INTRODUCTION

Recently, Bayesian learning of hidden Markov model (HMM) parameters has been proposed and adopted in a number of adaptive speech recognition applications [6, 1, 2, 3, 4]. In a conventional HMM-based Bayesian adaptation framework, HMM parameters of different speech units are usually assumed independent. Therefore, each model can only be adapted if the corresponding speech unit has been observed in the current adaptation data. Consequently, only after all units have been observed enough times, all of the HMM parameters can thus be effectively adapted. To enhance the efficiency and the effectiveness of the Bayes adaptive training, it is desirable to introduce some constraints on HMM parameters based on all possible sources of knowledge. Therefore all the model parameters can be adjusted at the same time in a consistent and systematic way even though some units are not seen in adaptation data. One possible way to achieve the above objective is to explicitly consider the correlation of HMM parameters corresponding to different speech units. However, it is too difficult to define a joint prior probability density function (PDF) for all sets of HMM parameters, if not impossible. A tractable case could be to assume all mean vectors are correlated and have a joint prior distribution [5]. In this paper, we restrict ourselves to this special case and extend our quasi-Bayes (QB) learning framework [2, 3, 4] to cope with the correlated continuous density HMMs (CDHMMs) with Gaussian mixture state observation densities.

Based on the theory of recursive Bayesian inference, the QB algorithm is designed to incrementally update the hyperparameters on the approximate posterior distribution and the CDHMM parameters simultaneously [3, 4]. By further introducing a simple forgetting mechanism [4] to adjust the contribution of previously observed sample utterances, the algorithm is truly adaptive in nature and capable of performing a full-scale on-line adaptive learning using only the current sample utterance. On the other hand, the QB framework is also flexible enough to include the batch or block mode learning as a special case.

Considering the difficulties of parameter updating and initial hyperparameters' estimation arisen from the introduction of correlation between different models, we propose, in this paper, a successive approximation algorithm based on pairwise correlations to update the mean vectors of CDHMMs as well as the corresponding hyperparameters. As an example, the method is applied to on-line speaker adaptation and its viability is confirmed in a series of comparative experiments using a 26-letter English alphabet vocabulary.

2. QUASI-BAYES LEARNING OF CORRELATED CDHMMS

Consider a collection of M CDHMMs $\Lambda = \{\lambda_q\}_{q=1,\cdots,M}$, where $\lambda_q = (\pi^{(q)}, A^{(q)}, \theta^{(q)})$ denotes the set of parameters of the q-th N-state CDHMM used to characterize the q-th speech unit, of which, $\pi^{(q)}$ is the initial state distribution, $A^{(q)} = [a_{ij}^{(q)}]$ is the transition probability matrix, and $\theta^{(q)}$ is the parameter vector composed of mixture parameters $\theta_i^{(q)} = \{\omega_{ik}^{(q)}, m_{ik}^{(q)}, \Sigma_{ik}^{(q)}\}$ for each state i with the state observation density being a mixture of multivariate Gaussian PDFs: $p(\mathbf{x}|\theta_i^{(q)}) = \sum_{k=1}^{K} \omega_{ik}^{(q)} \mathcal{N}(\mathbf{x}|m_{ik}^{(q)}, \Sigma_{ik}^{(q)})$, where the mixture coefficients $\omega_{ik}^{(q)}$'s satisfy the constraint $\sum_{k=1}^{K} \omega_{ik}^{(q)} = 1$, and $\mathcal{N}(\mathbf{x}|m_{ik}^{(q)}, \Sigma_{ik}^{(q)})$ is the k-th normal mixand with $m_{ik}^{(q)}$ being the D-dimensional mean vector and $\Sigma_{ik}^{(q)}$ being the $D \times D$ covariance matrix with its d-th diagonal element being $\sigma_{ik}^{(q)2}(d)$. For notational convenience, it is assumed that all the state observation PDFs have the same number of mixture components.

Let $\mathcal{X}_1^n = \{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_n\}$ be n independent sets of observation samples which are used to estimate the CDHMM

The first author would like to thank Drs. Y. Yamazaki and Y. Sagisaka of ATR-ITL for their support of this work.

parameters Λ. Our initial knowledge about Λ is assumed to be contained in a known joint *a priori* density $p(\Lambda)$. Let's assume the samples $\mathcal{X}_i$'s are given successively one by one, we can obtain a recursive expression for the *a posteriori* PDF of Λ, given $\mathcal{X}_1^n$, as

$$p(\Lambda|\mathcal{X}_1^n) = \frac{p(\mathcal{X}_n|\Lambda)\cdot p(\Lambda|\mathcal{X}_1^{n-1})}{\int_\Omega p(\mathcal{X}_n|\Lambda)\cdot p(\Lambda|\mathcal{X}_1^{n-1})d\Lambda}\ . \tag{1}$$

Starting the calculation of posterior PDF from $p(\Lambda)$, repeated use of the equation (1) produces the sequence of densities $p(\Lambda|\mathcal{X}_1^1)$, $p(\Lambda|\mathcal{X}_1^2)$, and so forth. This provides a basis of making formal recursive Bayesian inference of parameters Λ. However, there are some serious computational difficulties to directly implement this learning procedure [4]. Consequently, some approximations are needed in practice.

In this study, we only consider the case of CDHMMs in which the covariance matrices are specified. We define the parameter vector $\mathbf{m}$ to be the collection of the mean vectors of all the Gaussian mixture components of CDHMMs and denoted simply by an operator "vec" as $\mathbf{m} = vec\{m_{ik}^{(q)}\}$. We also define another operator "$block-diag$" to denote a block diagonal matrix, e.g., $\Xi = block-diag\{\Sigma_{ik}^{(q)}\}$, with each diagonal block element to be also a matrix, e.g., $\Sigma_{ik}^{(q)}$. Further denote $\lambda_q' = (\pi_i^{(q)}, a_{ij}^{(q)}, \omega_{ik}^{(q)})$. The initial prior PDF of Λ is assumed to be: $g(\Lambda) = g(\mathbf{m})\prod_{q=1}^{M} g(\lambda_q')$, where $g(\lambda_q')$ takes the special form of a matrix beta PDF with sets of positive hyperparameters of $\{\eta_i^{(q)}\}, \{\eta_{ij}^{(q)}\}, \{\nu_{ik}^{(q)}\}$ [1, 2], and $g(\mathbf{m}) = \mathcal{N}(\mathbf{m}|\mu, \mathbf{U})$ has a joint normal PDF with mean vector $\mu = vec\{\mu_{ik}^{(q)}\}$ and covariance matrix $\mathbf{U}$ [5]. This class of prior distributions actually constitutes a conjugate family of the complete-data density and is denoted as $\mathcal{P}$.

The quasi-Bayes procedure is, at each step of the recursive Bayes learning, to approximate the true posterior distribution $p(\Lambda|\mathcal{X}_1^n)$, by the "closest" tractable distribution $g(\Lambda|\varphi^{(n)})$ within the given class $\mathcal{P}$, under the criterion of both distributions having the same (local) mode. Here $\varphi^{(n)}$ denotes the updated hyperparameters after observing the samples $\mathcal{X}_n$. More specifically, consider at time instant n, we have a training set $\mathcal{X}_n = \{\mathbf{x}_n^{(q,r)}\}$ and our prior knowledge about Λ is approximated by $g(\Lambda|\varphi^{(n-1)})$. Here $\mathbf{x}_n^{(q,r)}$ denotes the rth training observation sequence of length $T_n^{(q,r)}$ associated with the q-th speech unit, and each unit has $W_q^{(n)}$ such observation sequences. Let $\mathcal{Y}_n = (\mathcal{X}_n, \mathcal{Z}_n)$ denote the associated complete-data and $\mathcal{Z}_n = \{\mathbf{s}_n^{(q,r)}, \mathbf{l}_n^{(q,r)}\}$ be corresponding missing-data, where $\mathbf{s}_n^{(q,r)}$ denotes the unobserved state sequence and $\mathbf{l}_n^{(q,r)}$ is the sequence of the unobserved mixture component labels corresponding to the observation sequence $\mathbf{x}_n^{(q,r)}$.

Given the set of observation sequences $\{\mathbf{x}_n^{(q,r)}\}$ and the above prior PDF $g(\Lambda|\varphi^{(n-1)})$, we can get the approximate MAP (maximum *a posteriori*) estimate $\Lambda^{(n)}$ of Λ by repeating following EM steps:

E-step: Compute

$$R(\Lambda|\Lambda^{(n-1+\frac{l-1}{L})}) = \kappa\cdot\log g(\Lambda|\varphi^{(n-1)}) + E[\log p(\mathcal{Y}_n|\Lambda)|\mathcal{X}_n, \Lambda^{(n-1+\frac{l-1}{L})}]\ , \tag{2}$$

where $0 < \kappa \leq 1$ is a forgetting factor and $\kappa = 1$ means no forgetting;

M-step: Choose

$$\Lambda^{(n-1+l/L)} = \underset{\Lambda}{\operatorname{argmax}}\, R(\Lambda|\Lambda^{(n-1+\frac{l-1}{L})}); \tag{3}$$

where $l = 1, 2, \cdots, L$ is the iteration index and L is the total number of iterations performed.

Hyperparameters $\varphi^{(n)}$ are updated at the last (actually Lth) iteration: $g(\Lambda|\varphi^{(n)}) \propto \exp\{R(\Lambda|\Lambda^{(n)})\}$.

For notational simplicity, from now on, we drop the related subscripts and/or superscripts which indicate the EM iteration index and global quasi-Bayes recursion index. The updating formulas for the CDHMM parameters $\{\pi_i^{(q)}, a_{ij}^{(q)}, \omega_{ik}^{(q)}\}$ remain the same as those in [4] while the mean vectors $\mathbf{m}$ is updated as follows:

$$\hat{\mathbf{m}} = \kappa\Xi(\kappa\Xi + \mathbf{UC})^{-1}\mu + \mathbf{U}(\kappa\Xi + \mathbf{CU})^{-1}\mathbf{C}\bar{\mathbf{X}} \tag{4}$$

where

$$\mathbf{C} = block-diag\{c_{ik}^{(q)}\cdot I_{D\times D}\} \tag{5}$$

$$\bar{\mathbf{X}} = vec\{\bar{\mathbf{x}}_{ik}^{(q)}\} \tag{6}$$

with

$$c_{ik}^{(q)} = \sum\nolimits_{r=1}^{W_q}\sum\nolimits_{t=1}^{T^{(q,r)}} \zeta_t^{(q,r)}(i,k) \tag{7}$$

$$\bar{\mathbf{x}}_{ik}^{(q)} = \sum\nolimits_{r=1}^{W_q}\sum\nolimits_{t=1}^{T^{(q,r)}} \zeta_t^{(q,r)}(i,k)\cdot\mathbf{x}_t^{(q,r)}/c_{ik}^{(q)} \tag{8}$$

$$\zeta_t^{(q,r)}(i,k) = \Pr(s_t^{(q,r)} = i, l_t^{(q,r)} = k|\mathbf{x}^{(q,r)}, \Lambda) \tag{9}$$

and $I_{D\times D}$ is an identity matrix.

At the last EM iteration, the hyperparameters $\{\eta_i^{(q)}, \eta_{ij}^{(q)}, \nu_{ik}^{(q)}\}$ are updated in the same way as those in [4], μ is updated as in (4), and $\mathbf{U}$ is updated as follows:

$$\hat{\mathbf{U}} = \mathbf{U}(\kappa\Xi + \mathbf{CU})^{-1}\Xi \tag{10}$$

Theoretically speaking, this completes the basic QB learning algorithm of CDHMMs with jointly correlated mean vectors. We also expect that this approximate recursive MAP estimate will converge asymptotically to its ML (maximum likelihood) batch counterpart as more and more adaptation data become available. However, in practice, it is very difficult to directly manipulate the updating formulas related to correlated mean vectors. The first difficulty comes from the estimation of the covariance matrix of the initial joint prior distribution of means due to the huge size of matrix. For example, in the above general formulation, covariance matrix $\mathbf{U}$ is of size $\mathcal{M}\times\mathcal{M}$ matrix ($\mathcal{M} = M\cdot N\cdot K\cdot D$). This means we need at least $\mathcal{M}+1$ sets of samples of mean vectors to get an estimation of a nonsingular covariance matrix $\mathbf{U}$ and this is usually impractical. The second difficulty lies in its computational complexity and memory requirement of algebraic manipulation involving such a huge-size matrix. Consequently, in practice, some simplifying assumptions should be attempted to make the algorithm useful. We provide one such solution in next section.

3. SUCCESSIVE APPROXIMATION BASED ON PAIRWISE CORRELATION

Suppose that we do not consider the correlation between different dimensional elements of the same mean vector or two different mean vectors. This can simplify the following discussion to a one-dimensional case. Further suppose that we only have the knowledge of pairwise correlations between different mean vectors instead of trying to exploit the joint correlation structure of all the mean vectors. So, every time, we only consider a pair of random variables $m_{ikd}^{(q)}$ and $m_{i'k'd}^{(q')}$. For notational simplicity, they are denoted respectively as $m_I(d)$ and $m_{I'}(d)$. We assume $m_I(d)$ and $m_{I'}(d)$ have a joint *a priori* normal PDF with means $\mu_I(d)$ and $\mu_{I'}(d)$, variances $u_I^2(d)$ and $u_{I'}^2(d)$, and covariance $\rho_{II'}(d) \cdot u_I(d) \cdot u_{I'}(d)$, where $\rho_{II'}(d)$ is the correlation coefficient. We pretend only $c_I = c_{ik}^{(q)}$ observations belonging to m_I are obtained and no observations for $m_{I'}$ are available. Given these observations, it can be shown that the joint posterior PDF of $m_I(d)$ and $m_{I'}(d)$ is still a normal one with the following hyperparameters:

$$\tilde{\mu}_I(d) = \mu_I(d) + \frac{c_I u_I^2(d)}{\kappa\sigma_I^2(d) + c_I u_I^2(d)}(\bar{x}_I(d) - \mu_I(d)) \quad (11)$$

$$\tilde{\mu}_{I'}(d) = \mu_{I'}(d) + \frac{c_I \rho_{II'}(d) u_I(d) u_{I'}(d)}{\kappa\sigma_I^2(d) + c_I u_I^2(d)}(\bar{x}_I(d) - \mu_I(d)) \quad (12)$$

$$\tilde{u}_I^2(d) = \frac{\sigma_I^2(d)}{\kappa\sigma_I^2(d) + c_I u_I^2(d)} u_I^2(d) \quad (13)$$

$$\tilde{u}_{I'}^2(d) = \frac{\kappa\sigma_I^2(d) + c_I u_I^2(d)(1 - \rho_{II'}^2(d))}{\kappa(\kappa\sigma_I^2(d) + c_I u_I^2(d))} u_{I'}^2(d) \quad (14)$$

$$\tilde{\rho}_{II'}(d) = \frac{\rho_{II'}(d)}{\sqrt{1 + \frac{c_I u_I^2(d)}{\kappa\sigma_I^2(d)}(1 - \rho_{II'}^2(d))}} \quad (15)$$

From the above results, we thus propose a successive approximation algorithm to update the equations (4) and (10). We then come up with the following on-line adaptation algorithm for correlated CDHMMs:

1. Estimate initial hyperparameters (details given in the next section). Set up (initial) top $\mathcal{K}$ prediction tables (explained in the following section).

2. Receive (an) utterance(s) to be recognized.

3. Do acoustic normalization/equalization as required.

4. Do recognition and record results.

5. Do supervised (if permitted) or unsupervised incremental adaptation as follows:

- in case of changeable top $\mathcal{K}$ tables, update them based on current correlation coefficients; otherwise, skip this step.
- do EM-iterations as follows:
  - initialize hyperparameters to be the latest history ones.
  - for those speech unit having observation data
    - update state transition matrices.
    - update mixture coefficients.
  - update mean vectors with successive approximation algorithm as follows:
    - reset temporary hyperparameters.
    - choose a mixture component "I" having observation data but not processed
      - * identify top $\mathcal{K}$ mixture components "I'"'s most correlated to mixture component I
      - * for each mixture component I', update its temporary hyperparameters as in equations (12), (14) and (15).
      - * update temporary hyperparameters for mixture component I as in equations (11) and (13)
    - if all the mixture components having observation data have been processed, go to next substep; otherwise, go back to previous substep.
    - update all mean vectors and exit the successive approximation algorithm.
- update all hyperparameters.

6. Go to Step 2.

4. IMPLEMENTATION ISSUES

4.1. Initial Hyperparameter Estimation

Apart from correlation coefficients, other initial hyperparameters can be estimated as in [4]. We use a modified *method of moment* to estimate the initial correlation coefficients $\rho_{II'}(d)$ as follows:

$$\frac{\sum_i c_I^{(i)}(m_I^{(i)}(d) - \bar{m}_I(d)) c_{I'}^{(i)}(m_{I'}^{(i)}(d) - \bar{m}_{I'}(d))}{\sqrt{\sum_i c_I^{(i)2}(m_I^{(i)}(d) - \bar{m}_I(d))^2} \cdot \sqrt{\sum_i c_{I'}^{(i)2}(m_{I'}^{(i)}(d) - \bar{m}_{I'}(d))^2}} \quad (16)$$

where $m_I^{(i)}$ is the i-th set of mean vectors, $c_I^{(i)}$ is the corresponding "EM count", and $\bar{m}_I$ is the average of $m_I^{(i)}$'s. In the following speaker adaptation (SA) experiments, we use speaker independent (SI) trained parameters to replace $\bar{m}_I$, and $m_I^{(i)}$ correspond to the parameters estimated from i-th speaker group (we used 16 speaker groups, see [2]).

4.2. Possible Constraints

For each mixture component having observation data, we only use its observed information to predict other $\mathcal{K}$ mixture components which have the highest top $\mathcal{K}$ values based on, among many possibilities, the following *between-component correlation measure*: $\bar{\rho}_{II'} = \frac{1}{D}\sum_{d=1}^{D} |\rho_{II'}(d)|$. In the following experiments, we only consider the correlation of mixture components between different speech units. Further constraints can also be applied to limit the correlated mixture components' domain based on some acoustic-phonetic knowledge (e.g., only consider the correlation between different speech units with a similar acoustic nature) and/or some data-driven clustering results. We will not further investigate these engineering issues here and leave them for future study.

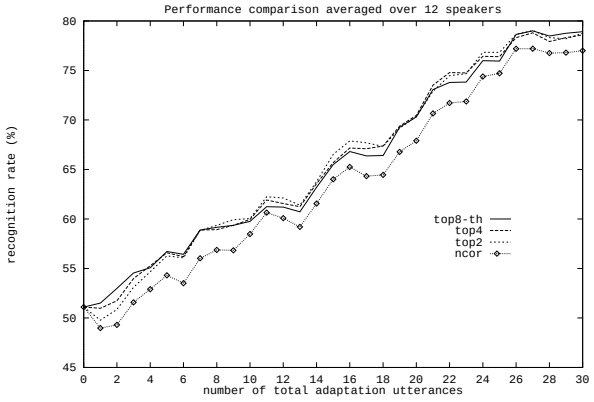

Figure 1: Performance comparison as a function of number of adaptation data (fast adaptation effect, $\kappa = 1$)

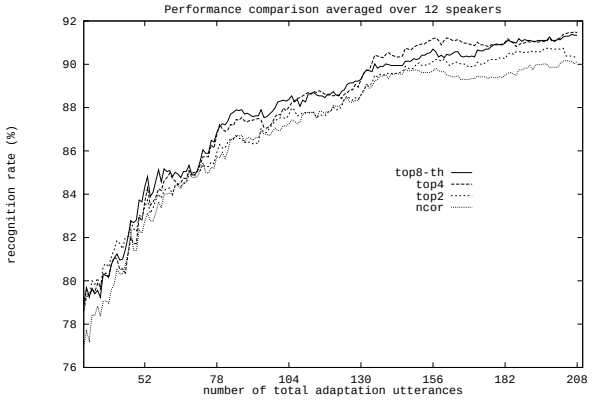

Figure 2: Performance comparison as a function of number of adaptation data (asymptotic convergence property, $\kappa = 1$)

5. SPEAKER ADAPTATION EXPERIMENTS

To examine its viability, the proposed algorithm is applied to on-line speaker adaptation. 26 letters of English alphabet are chosen as vocabulary. Two severely mismatched databases are used for evaluating the adaptation algorithm [2, 4]. For SI training and initial prior density estimation, the OGI ISOLET database produced by 150 speakers was used. For on-line SA training and testing, the TI46 isolated word corpus produced by 12 speakers was used. For each person and each letter, we divide equally those 16 tokens collected in 8 different sessions into two parts, one for adaptive training, another for testing. Throughout experiments, each letter is modeled by a left-to-right 5-state CDHMM with arbitrary state skipping and each state has 4 Gaussian mixture components with diagonal covariance matrix. Each feature vector consists of 12 LPC-derived cepstral coefficients and utterance-based cepstral mean subtraction (CMS) is applied for acoustic normalization.

Starting with a set of SI initial models, we present training tokens for each letter cyclically and perform utterance-based supervised on-line adaptation (OLA). After each OLA step, we test the recognizer on a separate testing set to measure the performance changes. We plot in Figures 1 and 2 the performance comparison of several OLA setups, averaged over 12 speakers, as a function of number of total adaptation tokens. In these figures, "ncor" stands for the experiment without considering correlation between mixture components. "top2" refer to the case of considering top 2 mixture components prediction and similar meaning for "top4". Apart from above meaning, "top8-th" also refer to the case of further applying a constraint of only considering the prediction of those mixture components whose correlation measures are above a threshold value. In these experiments, the top $\mathcal{K}$ prediction table is fixed and no forgetting mechanism is activated. Figure 1 shows the fast adaptation effects while Figure 2 checks the asymptotic property of the algorithm. The experimental results show that the proposed algorithm improves the OLA performance further by considering the correlation information and also has a good asymptotic convergence behavior.

6. CONCLUSION

In this paper, we extend our previously proposed on-line quasi-Bayes adaptive learning framework to handle the correlated CDHMM parameters. A successive approximation algorithm is proposed to implement the correlated mean vectors' updating. Its viability and efficacy are experimentally confirmed by an example of on-line speaker adaptation application. The same formulation can also be used to cope with varying channels, environments, and transducer mismatch problems in speech as well as speaker and other pattern recognition problems.

REFERENCES

[1] 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 Processing*, Vol. 2, pp.291-298, 1994.

[2] Q. Huo, C. Chan and C.-H. Lee, "Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 3, pp.334-345, 1995.

[3] Q. Huo, C. Chan and C.-H. Lee, "On-line adaptation of the SCHMM parameters based on the segmental quasi-Bayes learning for speech recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 4, pp.141-144, 1996.

[4] Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate," submitted to *IEEE Trans. on SAP*, 1995. See also a condensed version in *Proc. ICASSP-96* (Atlanta), May 1996.

[5] M. J. Lasry and R. M. Stern, "*A posteriori* estimation of correlated jointly Gaussian mean vectors," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. PAMI-6, pp.530-535, 1984.

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