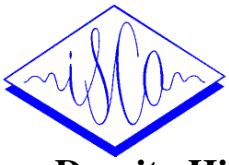

7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

Online Adaptation of Continuous Density Hidden Markov Models Based on Speaker Space Model Evolution

Dong Kook Kim and Nam Soo Kim

School of Electrical Engineering
Institute of New Media and Communications
Seoul National University, Seoul, Korea
{dkkim11, nkim}@snu.ac.kr

ABSTRACT

In this paper, we propose a new approach to online adaptation of continuous density hidden Markov model (CDHMM) based on speaker space model evolution. The speaker space model which characterizes the a priori knowledge of the training speakers is effectively described in terms of the latent variable model such as the factor analysis (FA) or probabilistic principal component analysis (PPCA). The latent variable models are employed to provide not only the speaker space model but also a joint prior distribution of CDHMM parameters, which can be directly applied to the maximum a posteriori (MAP) adaptation framework. We establish an online adaptation scheme based on the quasi-Bayes (QB) estimation technique which incrementally updates the hyperparameters of the speaker space model and the CDHMM parameters simultaneously. In a series of speaker adaptation experiments on the task of continuous digit recognition, we demonstrate that the proposed approach not only achieves a good performance for a small amount of adaptation data but also maintains a good asymptotic convergence property as the data size increases.

1. INTRODUCTION

Many online adaptation techniques for continuous-density hidden Markov model (CDHMM) have been studied to reduce the acoustic mismatches between the training and testing conditions [1]. These approaches make an automatic speech recognition (ASR) system capable of continuously adapting to the changing environments without waiting for a large set of adaptation data [1].

Huo and Lee [2] applied the quasi-Bayes (QB) learning framework to incrementally update both the CDHMM parameters and the hyperparameters through a prior evolution procedure. In addition, they extended the QB estimate to cope with the correlated CDHMM parameters in which all CDHMM mean vectors are correlated with a joint prior distribution [3]. More recently, another adaptation technique called the multiple-stream prior evolution and posterior pooling approach has been proposed to improve the ASR system performance with various amount of adaptation data [4]. Chien [5] presented an online transformation-based QB adaptation algorithm by adopting a simple transformation and assuming a specific prior probability density function (pdf) for these transformation parameters to improve the performance for both the sparse and sufficient adaptation data.

Recently, we proposed a rapid speaker adaptation technique based on the probabilistic principle component analysis (PPCA) [6] to extend the eigenvoice [7] approach such that it can be resolved into the maximum a posteriori (MAP) adaptation framework more easily [8]. The PPCA in which the probability density model is incorporated to the conventional PCA [9] finds the principal speaker space models based on the expectation maximization (EM) algorithm [10]. The PPCA model provides not only the speaker space models but also the a priori distribution of the model parameters, which can be directly applied to the MAP estimation scheme [8].

In this paper, we further extend the PPCA-based approach to the latent variable model such as factor analysis (FA) [11] to find the speaker space model, and propose a new approach to online adaptation of the CDHMM mean parameters based on the speaker space model evolution. The experimental results on incremental adaptation experiments show that the proposed approach not only achieves a good performance for a small amount of adaptation data but also guarantees a consistent estimate as the data size grows.

2. LATENT VARIABLE MODELS

Two examples of the latent variable model investigated here are FA and PPCA [6] [11]. Consider an observation data $\mathbf{y}$ of D-dimensional vector which is related to a latent variable $\mathbf{v}$ of dimension $P(\ll D)$ by

$$\mathbf{y} = \mathbf{W}\mathbf{v} + \bar{\mathbf{y}} + \epsilon \tag{1}$$

where $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_P]$ is a $D \times P$ matrix that represents the subspace of the observation data, $\bar{\mathbf{y}}$ is the mean of $\mathbf{y}$ and ϵ is a Gaussian random noise independent of $\mathbf{v}$. Traditionally, the latent variable is defined to be an independent Gaussian of unit variance, $p(\mathbf{v}) \sim \mathcal{N}(0, \mathbf{I}_P)$, where $\mathbf{I}_P$ is the $P \times P$ identity matrix. The main difference between the FA and PPCA lies in how the noise model is defined. In the FA model, the noise is characterized by a Gaussian with a diagonal covariance matrix $\mathbf{\Lambda}$ such that $p(\epsilon) \sim \mathcal{N}(0, \mathbf{\Lambda})$. On the other hand, the PPCA defines the noise covariance matrix to be isotropic, i.e., $\mathbf{\Lambda} = \sigma^2\mathbf{I}_D$, where $\mathbf{I}_D$ is the $D \times D$ identity matrix. Based on the above assumptions, the observation vector is also normally distributed as $p(\mathbf{y}) \sim \mathcal{N}(\bar{\mathbf{y}}, \mathbf{\Lambda} + \mathbf{W}\mathbf{W}^T)$. The pdf of $\mathbf{y}$ conditioned on $\mathbf{v}$ is given by

$$p(\mathbf{y}|\mathbf{v}) \sim \mathcal{N}(\mathbf{W}\mathbf{v} + \bar{\mathbf{y}}, \mathbf{\Lambda}) \, . \tag{2}$$

By means of the Bayes' theorem, the a posteriori distribution of the latent variable $\mathbf{v}$ given an observation vector $\mathbf{y}$ may be derived as

$$p(\mathbf{v}|\mathbf{y}) \sim \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^T\mathbf{\Lambda}^{-1}(\mathbf{y} - \bar{\mathbf{y}}), \mathbf{M}^{-1}) \tag{3}$$

where $\mathbf{M} = (\mathbf{I_P} + \mathbf{W}^T\mathbf{\Lambda}^{-1}\mathbf{W})^{-1}$. Given an observation vector sequence $\mathbf{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_T\}$, the latent variable model estimates the latent variable sequence $\mathbf{V} = \{\mathbf{v}_1, \cdots, \mathbf{v}_T\}$ and finds the optimal model parameters $\hat{\phi} = \{\hat{\mathbf{W}}, \hat{\bar{\mathbf{y}}}, \hat{\mathbf{\Lambda}}\}$ according to the maximum likelihood (ML) criterion. Since, however, the latent variables $\{\mathbf{v}_t\}$ are considered to be hidden, the EM algorithm which iteratively updates the parameter values is applied. Let $\phi^{(n)}$ be the parameter values obtained at the nth iteration. Then, the new parameter values $\hat{\phi}$ are obtained by

$$\hat{\phi} \quad = \quad \underset{\phi}{argmax}\ E\left[\log p(\mathbf{Y}, \mathbf{V}|\phi)|\mathbf{Y}, \phi^{(n)}\right] . \qquad (4)$$

3. ONLINE ADAPTATION BASED ON SPEAKER SPACE MODEL EVOLUTION

3.1. Speaker space model

Consider an N-state CDHMM with K mixture components, $\lambda = \{\lambda_j\} = \{w_{jk}, \boldsymbol{\mu}_{jk}, \Sigma_{jk}\}, j = 1, \cdots, N, k = 1, \cdots, K$. The state observation pdf of an observation vector $\mathbf{x}$ is defined to be a mixture of multivariate Gaussians

$$p(\mathbf{x}|\lambda_j) = \sum_{k=1}^{K} w_{jk}\mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}_{jk}, \Sigma_{jk}\right) \qquad (5)$$

where w_{jk} is the weight for the mixture component k in state j, $\boldsymbol{\mu}_{jk}$ is the d-dimensional mean vector and Σ_{jk} is the $d \times d$ covariance matrix.

Let $\{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_R\}$ be a set of R well-trained speaker dependent (SD) models, which can be obtained from the given training database. Here, $\boldsymbol{\mu}_r = [\boldsymbol{\mu}_{r,00}^T, \cdots, \boldsymbol{\mu}_{r,NK}^T]^T$ is the supervector of dimension D that corresponds to all the Gaussian mean vectors from the rth speaker model. Specifically, $\boldsymbol{\mu}_{r,jk}$ represents the mean vector of the kth Gaussian in the jth state of the rth speaker CDHMM. We assume that a set of SD models, $\{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_R\}$ are generated by a latent variable model with parameters $\phi = \{\boldsymbol{W}, \bar{\boldsymbol{\mu}}, \mathbf{\Lambda}\}$ such that

$$\boldsymbol{\mu}_r \quad = \quad \sum_{i=1}^{P} \mathbf{w}_i v_i + \bar{\boldsymbol{\mu}} + \epsilon \qquad (6)$$

where $\bar{\boldsymbol{\mu}}$ is the mean of the supervectors, $\mathbf{w}_i$ is a vector denoting the ith column of $\mathbf{W}$ and $\mathbf{v} = [v_1, \cdots, v_P]^T$ is the weight vector representing a point within the speaker space which is spanned by the P column vectors of $\mathbf{W}$. Then, the speaker space model parameters, ϕ can be estimated using the iterative EM algorithm. Once the set of parameters, ϕ is given, we can construct a prior pdf of $\boldsymbol{\mu}$ such that

$$g(\boldsymbol{\mu}|\phi) \sim \mathcal{N}(\bar{\boldsymbol{\mu}}, \mathbf{\Lambda} + \mathbf{W}\mathbf{W}^T) . \qquad (7)$$

It is noted that the latent variable model provides the information of correlation among different speech units as well as the prior pdf associated with CDHMM mean parameters.

Given the latent variable $\mathbf{v}$, the speaker model $\boldsymbol{\mu}$ does not appear as a single point in the speaker space but a random variable with the pdf $p(\boldsymbol{\mu}|\mathbf{v})$ due to the noise process. Estimating the latent variable $\mathbf{v}$ generates the a priori pdf of the speaker model instead of producing directly the adapted model. This prior pdf which represents the estimated prior knowledge for a specific speaker enables us to employ the MAP-based speaker adaptation scheme.

3.2. QB learning for speaker space model

In this subsection, we briefly review the basic concept and formulation of QB learning [1]-[5]. Let $\mathcal{X}^n = \{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_n\}$ be n independent sets of observation data which are incrementally collected to update the CDHMM parameters, λ. A recursive expression for the a posteriori pdf of λ is given by

$$p(\lambda|\mathcal{X}^n) = \frac{p(\mathcal{X}_n|\lambda) \cdot p(\lambda|\mathcal{X}^{n-1})}{\int p(\mathcal{X}_n|\lambda) \cdot p(\lambda|\mathcal{X}^{n-1}) d\lambda} . \qquad (8)$$

This provides a basis for making a recursive Bayesian estimate of the given parameter λ. However, the implementation of this type of recursive Bayesian estimation technique on a CDHMM has been found very difficult. To alleviate this problem, an approach called the QB learning was proposed in [2]. The QB procedure, at each step of the recursive Bayes learning, approximates the true posterior density $p(\lambda|\mathcal{X}^n)$ by the closest tractable parametric prior density $g(\lambda|\phi^{(n)})$ under the criterion that both densities should have the same mode. Here, $\phi^{(n)}$ denotes the updated hyperparameters after observing $\mathcal{X}_n$.

Let us assume that at time instant n, we are given a set of observation vectors $\mathcal{X}_n = \{\mathbf{x}_1^{(n)}, \cdots, \mathbf{x}_{T_n}^{(n)}\}$ and the approximate prior pdf $g(\lambda|\phi^{(n-1)})$. Since we assume that the means contained in λ is generated through a model given by (6), which has a hidden variable $\mathbf{v}$ with the hyperparameter $\phi^{(n-1)}$, the complete-data likelihood for λ can be easily defined. Let $(\mathcal{X}_n, \mathbf{S}_n, \mathbf{L}_n)$ denote the complete-data for $\mathcal{X}_n$ in which $\mathbf{S}_n = \{s_t^{(n)}\}$ represents the state sequence and $\mathbf{L}_n = \{l_t^{(n)}\}$ is the mixture component sequence. We can obtain an approximate MAP estimate $\lambda^{(n)}$ of λ by repeating the following EM steps:
E-step: Compute the auxiliary function

$$\begin{aligned} R(\lambda|\hat{\lambda}_{m-1}) = \ & E\ [\ \log p(\mathcal{X}_n, \mathbf{S}_n, \mathbf{L}_n \mid \lambda) \\ & +\ \log g(\lambda, \mathbf{v} \mid \phi^{(n-1)}) \mid \mathcal{X}_n, \hat{\lambda}_{m-1}] . \end{aligned} \qquad (9)$$

M-step: Choose

$$\hat{\lambda}_m = \underset{\lambda}{argmax}\ R(\lambda|\hat{\lambda}_{m-1}) \qquad (10)$$

where $m = 1, \cdots, M$ and M denotes the total number of iterations performed with $\hat{\lambda}_0 = \lambda^{(n-1)}$. By repeating the above EM iteration, we can get a series of approximate pdf $g(\lambda|\phi^{(n)})$ whose mode is approaching the mode of the true posterior pdf $p(\lambda|\mathcal{X}_n)$. At the last EM iteration, the set of hyperparameters $\phi^{(n)}$ is computed to satisfy

$$g(\lambda|\phi^{(n)}) \propto \exp\left\{R(\lambda|\hat{\lambda}_{M-1})\right\} . \qquad (11)$$

Finally the CDHMM parameters $\lambda^{(n)}$ are updated by taking the mode of $g(\lambda|\phi^{(n)})$.

3.3. Speaker space model evolution

In this paper, we consider only the case of prior evolution for the mean vectors of CDHMM in which mixture weights and covariance matrices are fixed during adaptation. Under the specification of the speaker space model for $\boldsymbol{\mu}$ in (6), the auxiliary function in the expectation step can be rewritten as

$$\begin{aligned} R(\lambda|\hat{\lambda}_{m-1}) = \ & \sum_{\mathbf{S}_n} \sum_{\mathbf{L}_n} p(\mathbf{S}_n, \mathbf{L}_n|\mathcal{X}_n, \hat{\lambda}_{m-1}) \cdot \\ & \log p(\mathcal{X}_n, \mathbf{S}_n, \mathbf{L}_n|\lambda) + E\left[\log p(\lambda|\mathbf{v}, \phi^{(n-1)}) p(\mathbf{v}) \mid \hat{\lambda}_{m-1}\right] . \end{aligned} \qquad (12)$$

Based upon (2) and (5), it is not difficult to derive

$$
\begin{aligned}
&R(\lambda|\hat{\lambda}_{m-1}) \\
&= \sum_{t=1}^{T_n}\sum_{j=1}^{N}\sum_{k=1}^{K}\gamma_t(j,k)\left[-\frac{1}{2}(\mathbf{x}_t^{(n)} - \boldsymbol{\mu}_{jk})^T\Sigma_{jk}^{-1}(\mathbf{x}_t^{(n)} - \boldsymbol{\mu}_{jk})\right] \\
&\quad + E\,[-\frac{1}{2}(\boldsymbol{\mu} - \mathbf{W}^{(n-1)}\mathbf{v} - \bar{\boldsymbol{\mu}}^{(n-1)})^T\boldsymbol{\Lambda}^{-1,(n-1)} \\
&\quad\quad \cdot (\,\boldsymbol{\mu} - \mathbf{W}^{(n-1)}\mathbf{v} - \bar{\boldsymbol{\mu}}^{(n-1)})\mid \hat{\lambda}_{m-1}\,]
\end{aligned}
\tag{13}
$$

where $\gamma_t(j,k) = P(s_t^{(n)} = j, l_t^{(n)} = k|\mathcal{X}_n, \hat{\lambda}_{m-1})$ is the posterior probability of being in state j and mixture component k at time t given the observation sequence $\mathcal{X}_n$. From (11), $g(\lambda|\phi^{(n)})$ can be expressed as follows [12]:

$$
g(\lambda|\phi^{(n)}) \propto C\cdot\exp\left\{-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{m})^T\boldsymbol{\Phi}^{-1}(\boldsymbol{\mu}-\mathbf{m})\right\} \tag{14}
$$

where C is a normalization constant and

$$
\begin{aligned}
\mathbf{m} &= \boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1,(n-1)}(\mathbf{W}^{(n-1)}\hat{\mathbf{v}} + \bar{\boldsymbol{\mu}}^{(n-1)}) + \boldsymbol{\Phi}\mathbf{C}\mathbf{S}^{-1}\boldsymbol{\mu}_{ML} \\
\boldsymbol{\Phi} &= (\boldsymbol{\Lambda}^{-1,(n-1)} + \mathbf{C}\mathbf{S}^{-1})^{-1}
\end{aligned}
\tag{15}
$$

with $\mathbf{S} = diag(\Sigma_{11},\cdots,\Sigma_{NK})$ and $\mathbf{C} = diag(c_{11},\cdots,c_{NK})$ in which $c_{jk} = \sum_t \gamma_t(j,k)$ is the count of the kth Gaussian in the jth state observed along the adaptation data used. Furthermore,

$$
\begin{aligned}
\hat{\mathbf{v}} &= E\left[\mathbf{v}|\hat{\lambda}_{M-1}\right] \\
&= (\mathbf{I} + \mathbf{W}^{T,(n-1)}\boldsymbol{\Lambda}^{-1,(n-1)}\mathbf{W}^{(n-1)})^{-1}\cdot \\
&\quad \mathbf{W}^{T,(n-1)}\boldsymbol{\Lambda}^{-1,(n-1)}(\hat{\boldsymbol{\mu}}_{M-1} - \bar{\boldsymbol{\mu}}^{(n-1)})
\end{aligned}
\tag{16}
$$

and $\boldsymbol{\mu}_{ML} = [(\boldsymbol{\mu}_{11})_{ML}^T,\cdots,(\boldsymbol{\mu}_{NK})_{ML}^T]^T$ where $(\boldsymbol{\mu}_{jk})_{ML} = (\sum_t \gamma_t(j,k)\mathbf{x}_t^{(n)})/c_{jk}$.

(14) belongs to the same distribution family as $g(\boldsymbol{\mu}|\phi)$ in (7) with $\bar{\boldsymbol{\mu}} = \mathbf{m}$ and $\boldsymbol{\Lambda} + \mathbf{W}\mathbf{W}^{\mathbf{T}} = \boldsymbol{\Phi}$. Here we assume that the hyperparameter $\mathbf{W}$ does not evolve (i.e., fixed during prior evolution) and the prior evolution for $\boldsymbol{\Lambda}$ is approximated by $\boldsymbol{\Phi}$. As a result, $g(\lambda|\phi^{(n)})$ can be denoted with the hyperparameters $\phi^{(n)}$ as follows:

$$
\begin{aligned}
\bar{\boldsymbol{\mu}}^{(n)} &= \mathbf{m} \\
\boldsymbol{\Lambda}^{(n)} &= \boldsymbol{\Phi} \\
\mathbf{W}^{(n)} &= \mathbf{W}^{(n-1)}
\end{aligned}
\tag{17}
$$

such that the speaker space model is evolving accordingly. For the prior evolution of the PPCA model, similar results are obtained, $\bar{\boldsymbol{\mu}}^{(n)} = \mathbf{m}$ and $\sigma^{2,(n)} = trace\{\boldsymbol{\Phi}\}/D$ with $\boldsymbol{\Lambda}^{(n-1)} = \sigma^{2,(n-1)}\mathbf{I}$. After completing the speaker space model evolution procedure, the adapted CDHMM mean vectors $\boldsymbol{\mu}^{(n)}$ are obtained by just taking the mode of the evolved prior pdf as follows:

$$
\boldsymbol{\mu}^{(n)} = \mathbf{m}\,. \tag{18}
$$

It is noted that the update parameter $\boldsymbol{\mu}^{(n)}$ is obtained by simply interpolating the ML estimate of the new data with the prior speaker model estimated within the speaker space. The speaker space model evolution and the updated mean parameter by the latent variable model are illustrated in Fig. 1.

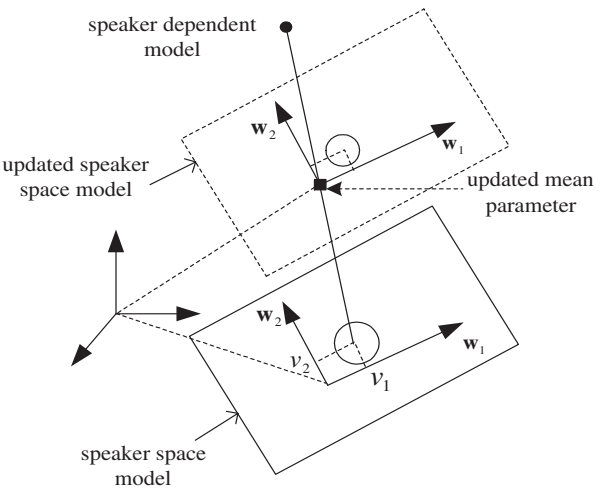

Figure 1: Speaker space model evolution by latent variable model with $D = 3$ and $P = 2$.

4. EXPERIMENTS AND RESULTS

4.1. Experimental setup

Performance of the proposed online adaptation approach was evaluated with a number of supervised adaptation experiments on the task of continuous Korean digit recognition. All the training and test data used for building the baseline recognition system were recorded in a quite environment. Utterances from 105 speakers constructed the training data and those from the other 35 speakers were used for evaluation. Each speaker contributed 30∼40 sentences consisting of 3∼7 digits. Each digit was modeled by a seven-state left-to-right HMM without skips and the 3 silence types were modeled by a one-state HMM. We trained the CDHMM parameters by varying the number of Gaussian components in each state from one to two. The speech signal was sampled at 8 kHz and segmented into 30 ms frame at every 10 ms with 20 ms overlap. Each frame was parameterized by a 24-dimensional feature vector consisting of 12 mel-frequency cepstral coefficients and their first-order time derivatives. In the recognition experiments, we drew $1 \sim 10$ sentences ($1 \sim 20$ sec) from each target speaker for adaptation, and performed the recognition test on the remaining sentences. All adaptation procedures were performed in a supervised manner using an exact transcription of each data. The speaker independent (SI) system with a single Gaussian and two mixture Gaussians produced 87.58 % and 89.6 % of word recognition rates, respectively.

To obtain the speaker space model by latent variable models, we first trained a set of SI models over the speech from all the training speakers. To obtain the 105 SD HMM's, the ML training approach was applied in the single Gaussian mixture system while the MAP-based adaptation algorithm was used in the two mixture Gaussian system for each training speaker. We extracted a supervector by augmenting all the mean vectors of each SD model. Consequently, the mean supervectors are of dimension $D = \{11\times7\times24+3\times24\}\times\{1,2\} = \{1920, 3840\}$. Before applying the latent variable model techniques, the supervectors were normalized by their standard deviation to prevent the variables with large absolute value from dominating the analysis [7] [9]. We obtained the estimates of the parameters for the latent variable model, $\{\bar{\boldsymbol{\mu}}, \mathbf{W}, \boldsymbol{\Lambda}\}$ with dimension $P = 10$ for each mean supervector using the EM algorithm.

4.2. Batch adaptation results

First, we performed supervised batch speaker adaptation experiments by using four different methods, i.e., 1) conventional MAP adaptation method, 2) eigenvoice method, 3) PPCA-based adaptation method, 4) FA-based adaptation method. Tables I and II show the performance of the conventional MAP adaptation method, the eigenvoice, PPCA and FA-based adaptation techniques for the single Gaussian and two mixture Gaussian HMM's, respectively. Word recognition rates are displayed against the number of adaptation sentences. Here "EV" represents the eigenvoice method and "PPCA" and "FA" denote the PPCA and FA methods, respectively.

The eigenvoice approach performed very well for a little adaptation data size, but they did not improve the performance as more data became available. On the other hand, the PPCA and FA-based approaches possess both the rapid and consistent adaptation properties, and as a result they can efficiently perform speaker adaptation not only for very little adaptation data but also when a large amount of data is available.

Table I
Word recognition rate (%) for batch adaptation experiments in a single Gaussian system.

no. of sent.	1	2	4	6	8	10
MAP	87.48	88.04	88.47	88.89	89.47	89.97
EV	88.62	89.91	90.26	90.28	90.26	90.24
PPCA	89.76	90.20	90.22	90.53	90.63	90.95
FA	89.76	90.28	90.43	90.76	90.97	91.38

Table II
Word recognition rate (%) for batch adaptation experiments in two mixture Gaussian system.

no. of sent.	1	2	4	6	8	10
MAP	89.52	89.8	89.97	90.43	90.88	91.28
EV	91.05	91.13	91.40	91.51	91.63	91.71
PPCA	91.01	91.44	91.90	92.05	92.36	92.79
FA	90.53	91.13	91.82	92.03	92.25	92.71

4.3. Online adaptation results

Fig. 2 shows the performance of the PPCA and FA-based online adaptation techniques with various amount of adaptation data. For online adaptation, the parameters were updated for each adaptation sentence. From Fig. 2, we can find that the online version of the proposed approach achieves a similar rapid adaptation performance as that of the batch PPCA and FA approaches for a small amount of adaptation data, while it maintains a good asymptotic convergence property as the data size grows. The experimental results show that the online- and batch-mode adaptation approaches which are based on the latent variable models perform equally well.

5. CONCLUSIONS

We have presented techniques to incrementally adapt the CDHMM parameters based on speaker space model evolution. We have applied the latent variable model to find the speaker space model associated with the mean vectors as well as to obtain their prior pdfs. We have extended the recursive QB learning to online speaker adaptation using the speaker space model. From the results of a set of online speaker adaptation experiments, we can conclude that the proposed approach is effective in speaker adaptation both for sparse and sufficient data.

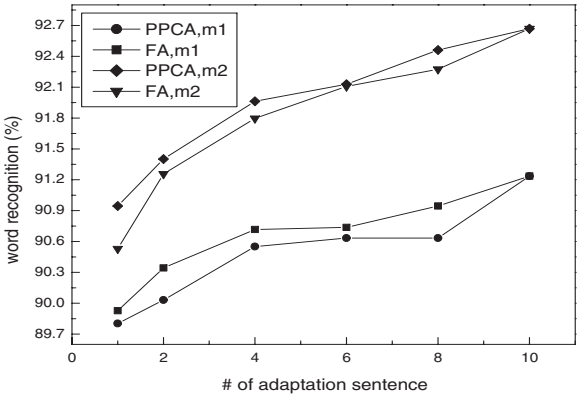

Figure 2: Word recognition rate for online adaptation experiments in a single Gaussian (m1) and two mixture Gaussian system (m2).

6. REFERENCES

[1] C.-H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition", Proc. of IEEE, vol. 88, pp.1241-1269, Aug. 2000.

[2] Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate," IEEE Trans. Speech and Audio Proc., vol. 5, pp. 161-172, Mar. 1997.

[3] ——, "On-line adaptive learning of the correlated continuous density hidden Markov models for speech recognition," IEEE Trans. Speech and Audio Proc., vol. 6, pp. 386-397, July 1998.

[4] Q. Huo and B. Ma, "Online adaptive learning of continuous-density hidden Markov models based on multiple-stream prior evolution and posterior pooling," IEEE Trans. Speech and Audio Proc., vol. 9, pp. 388-398, May 2001.

[5] J.-T. Chien, "On-line hierarchical transformation of hidden Markov models for speech recognition," IEEE Trans. Speech and Audio Proc., vol. 7, pp. 656-667, Nov. 1999.

[6] M. Tipping and C. Bishop, "Mixtures of probabilistic principal comonent analysers," Neural Computation, vol. 11, no. 2, pp. 443-482, 1999.

[7] R. Kuhn, J.-C. Junqua, P. Nguyen and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space", IEEE Trans. Speech and Audio Proc., vol. 8, pp. 695-707, Nov. 2000.

[8] D. K. Kim and N. S. Kim, "Baysian speaker adaptation based on probabilistic principal component analysis", Proc. of ICSLP, pp. 734-737, 2000.

[9] I. T. Jolliffe, Principal Component Analysis, Springer-Verlag, 1986.

[10] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, vol. 39, pp. 1-38, 1977.

[11] D. Rubin and D. Thayer, "EM algorithms for factor analysis," Psychometrika, vol. 47, pp. 69-76, 1982.

[12] G. Zavaliagkos, "Maximum a posteriori adaptation techniques for speech recognition," Ph.D. dissertation, Northeastern Univ., Boston, MA, 1995.