On-Line Adaptive Learning of the Continuous Density Hidden Markov Model Based on Approximate Recursive Bayes Estimate

Qiang Huo, *Member, IEEE,* and Chin-Hui Lee, *Fellow, IEEE*

***Abstract*—We present a framework of quasi-Bayes (QB) learning of the parameters of the continuous density hidden Markov model (CDHMM) with Gaussian mixture state observation densities. The QB formulation is based on the theory of recursive Bayesian inference. The QB algorithm is designed to incrementally update the hyperparameters of the approximate posterior distribution and the CDHMM parameters simultaneously. By further introducing a simple forgetting mechanism to adjust the contribution of previously observed sample utterances, the algorithm is adaptive in nature and capable of performing an on-line adaptive learning using only the current sample utterance. It can, thus, be used to cope with the time-varying nature of some acoustic and environmental variabilities, including mismatches caused by changing speakers, channels, and transducers. As an example, the QB learning framework 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.**

***Index Terms*— Recursive Bayesian estimation, incremental maximum likelihood estimation, hidden Markov model, EM algorithm, automatic speech recognition, speaker adaptation**

I. Introduction

CLASSICAL parameter estimation methods of hidden Markov model (HMM), such as maximum likelihood (ML) [4], [3], [24], [17] and maximum *a posteriori* (MAP) [21], [12], [14], generally imply batch algorithms that require processing the available data as a whole. In a variety of speech recognition applications, it is desirable to process the data sequentially. For example, in many speech recognition systems, there usually exists a performance gap between the recognition accuracies on training and on testing data. One major reason lies in the possible mismatch between the underlying acoustic characteristics associated with the training and testing conditions. This mismatch may arise from inter- and intraspeaker variabilities, transducer, channel and other environmental variabilities, and many other phonetic and linguistic effects due to a task mismatch problem. To bridge this performance gap, one possible solution is to design a speech recognition system that is robust to the above types of acoustic mismatch, and this has been a long standing objective of many researchers over the past 20 years. Another way to reduce the possible acoustic mismatch between the training and testing conditions is to adopt the so-called *adaptive learning* approach. The scenario is like this: starting from a pretrained (e.g., speaker and/or task independent) speech recognition system, for a new user (or a group of users) to use the system for a specific task, a small amount of adaptation data is collected from the user. These data are used to construct a speaker adaptive system for the speaker in the particular environment for that specific application. By doing so, the mismatch between training and testing can generally be reduced. The most fascinating adaptation scheme with a practical value is the so-called on-line (or incremental, sequential) adaptation. This scheme makes the recognition system capable of continuously adapting to the new adaptation data (possibly derived from actual test utterances) without the requirement of storing a large set of previously used training data. It is this kind of approach that this paper focuses on.

The advantage of a sequential algorithm over a batch algorithm is not necessarily in the final result, but in computational efficiency, reduced storage requirements, and the fact that an outcome may be provided without having to wait for all the data to be processed. Moreover, the parameters of interest are sometimes subject to changes, e.g., they are time varying just like abovementioned acoustic mismatch problem frequently encountered in real speech recognition applications. In such cases, different data segments often correspond to different parameter values. Processing of all the available data jointly is no longer desirable, even if we can afford the computational load of the batch algorithm. To alleviate such problems, a sequential algorithm can be designed to adaptively track the varying parameters.

Recently, Bayesian adaptive learning of HMM parameters has been proposed and adopted in a number of speech recognition applications. A theoretical framework of Bayesian learning was first proposed by Lee *et al.* [21] for estimating the mean and covariance matrix parameters of a continuous density HMM (CDHMM) with a multivariate Gaussian state observation density. It was then extended to handle all the parameters of a CDHMM with Gaussian mixture state observation densities (e.g., [12]) as well as the parameters of discrete HMM's (DHMM's) and semicontinuous HMM's (SCHMM's, also called tied-mixture HMM's) (e.g., [14]). It was shown that, for HMM-based speech recognition applications, the

Manuscript received September 28, 1995; revised September 15, 1996. The associate editor coordinting the review of this manuscript and approving it for publication was Prof. Joseph Picone.

Q. Huo is with ATR Interpreting Telecommunications Research Laboratories, Kyoto 619-02, Japan (e-mail: qhuo@itl.atr.co.jp).

C.-H. Lee is with Multimedia Communications Research Laboratory, Bell Laboratories, Murray Hill, NJ 07974 USA (e-mail: chl@research.bell-labs.com).

Publisher Item Identifier S 1063-6676(97)01900-7.

MAP framework provides an effective way for combining adaptation data and the prior knowledge, and then creating a set of adaptive HMM's to cope with the new acoustic conditions in the test data. The prior knowledge, which is embodied in a set of seed HMM's as well as in the assumed distributions of the model parameters being adapted, is made use of to mitigate the effect of adaptation data shortage to improve the system robustness. This approach works in a *batch* adaptation mode using a history of all the adaptation data. It can also be modified to work in a more attractive *incremental* adaptation mode. A related study was conducted by Matsuoka and Lee [28] in which they used the segmental MAP algorithm to perform the so-called on-line adaptation. Due to its missing mechanism of updating the hyperparameters of the prior and/or posterior distribution incrementally, all the previously seen adaptation data need to be stored. A full-scale *on-line* Bayesian adaptation approach should be able to update both the hyperparameters of the prior and/or posterior distributions and the HMM parameters themselves simultaneously upon the presentation of the latest adaptation data. One such approach for adapting the mixture coefficients of SCHMM parameters was recently developed in [14] and [15]. In this study, we expand the above work and investigate the incremental estimation of all of the CDHMM parameters. The formulation given here can be straightforwardly extended to the DHMM and SCHMM cases.

A block diagram of the proposed on-line Bayesian adaptive training of HMM's is shown in Fig. 1. Given a new block of input speech, feature extraction (usually spectral analysis) is first performed to derive the feature vector sequences used to characterize the speech input. It is followed by some kind of acoustic normalization to reduce the possible mismatch in the feature vector space. The processed feature vector sequences are then recognized based on the current set of HMM's. After the recognition of the current block of utterances, the HMM's and the posterior distributions of the related speech units are adapted and the updated models are used to recognize future input utterance(s). The adaptation algorithm usually requires some form of supervision in terms of the word (or phone) transcription of the speech utterances. Such a transcription can be provided either by a human transcriber or by the correction made by the user on the recognized output during actual usage. This adaptation scheme is often called *supervised* adaptation. On the other hand, the supervision information can also be derived directly from the recognition results and this is often referred to as *unsupervised* adaptation. For real-world applications, the unsupervised mode is usually more realistic and desirable. For the acoustic normalization/equalization module shown in Fig. 1, many existing techniques can be applied. They include, for example, the popular cepstral mean subtraction algorithm [2], different cepstral normalization methods (e.g., CDCN) discussed in [1], ML-based feature space stochastic matching methods [7], [41], [33], signal conditioning techniques [30], [31], etc. Acoustic normalization could even be integrated into the feature extraction stage, e.g., speaker normalization via vocal tract length normalization using frequency warping [39], [11], [22]. Encouraging results have also been demonstrated in combined acoustic normalization and model adaptation based on a small amount of calibration data [41], [42].

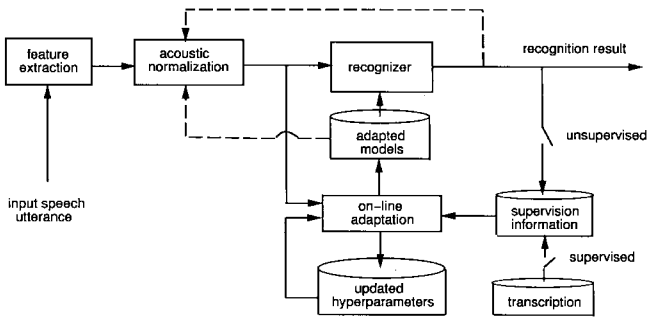

Fig. 1. Block diagram of on-line Bayesian adaptation of HMM's.

The rest of the paper is organized as follows. A brief introduction of the concept of recursive Bayesian inference for CDHMM is given in Section II. The difficulty of directly applying the recursive scheme is also illustrated. The formulation of approximate quasi-Bayes estimation for incremental CDHMM training is proposed in Section III. Some important implementation issues are discussed in Section IV. In Section V, a series of experimental results along with discussions and analyzes for an incremental speaker adaptation application are reported. Finally, we summarize our findings in Section VI.

II. Incremental Bayes Learning: Method and Difficulty

Consider an N-state CDHMM with parameter vector $\lambda = (\pi, A, \theta)$, where $\pi = [\pi_1, \pi_2, \cdots, \pi_N]^t$ is the initial state probability vector, $A = [a_{ij}]$, $i, j = 1, 2, \cdots, N$, is the transition probability matrix, and θ is the parameter vector composed of mixture parameters $\theta_i = \{\omega_{ik}, m_{ik}, r_{ik}\}$, $k = 1, 2, \cdots, K$ for each state i. The state observation probability density function (pdf) is assumed to be a mixture of multivariate Gaussian pdf's

$$p(\mathbf{x} \,|\, \theta_i) = \sum_{k=1}^{K} \omega_{ik} f_{ik}(\mathbf{x}) = \sum_{k=1}^{K} \omega_{ik} \mathcal{N}(\mathbf{x} \,|\, m_{ik}, r_{ik}) \qquad (1)$$

where the mixture coefficients ω_{ik}'s satisfy the constraint $\sum_{k=1}^{K} \omega_{ik} = 1$, and $\mathcal{N}(\mathbf{x} \,|\, m_{ik}, r_{ik})$ is the kth normal mixand denoted by

$$\mathcal{N}(\mathbf{x} \,|\, m_{ik}, r_{ik}) \propto |r_{ik}|^{1/2} \exp\left[-\frac{1}{2}(\mathbf{x} - m_{ik})^t r_{ik} (\mathbf{x} - m_{ik})\right] \qquad (2)$$

with m_{ik} being the D-dimensional mean vector and r_{ik} being the $D \times D$ precision (inverse covariance) matrix. Here, "$\propto$" denotes proportionality and $|r|$ denotes the determinant of the matrix r. Note that for notational convenience, it is assumed that the observation pdf's of all the states have the same number of mixture components. Whenever possible, in this paper, we try to use the same notations as in [12].

Let $\mathcal{X}^n = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n\}$ be n independent observation samples which are used to estimate the CDHMM parameters λ. Our initial knowledge about λ is assumed to be contained in a known *a priori* density $p(\lambda)$. A formal Bayesian inference

of λ is based on the following *a posteriori* density

$$p(\lambda \mid \mathcal{X}^n) = \frac{p(\mathcal{X}^n \mid \lambda) \cdot p(\lambda)}{\int_\Omega p(\mathcal{X}^n \mid \lambda) \cdot p(\lambda) d\lambda} \tag{3}$$

where Ω denotes an admissible region of the parameter space. The classical MAP solution of λ can be obtained by using the expectation maximization (EM) algorithm [9] substantiated in [12], which is an iterative algorithm working in batch mode to find a local maximum of the posterior pdf $p(\lambda \mid \mathcal{X}^n)$.

Assume the training/adaptation samples $\mathbf{X}_i$'s are presented successively. Applying the Bayes theorem, we obtain a recursive expression for the *a posteriori* pdf of λ, given $\mathcal{X}^n$, as

$$\begin{aligned} p(\lambda \mid \mathcal{X}^n) &= \frac{p(\mathbf{X}_n \mid \mathcal{X}^{n-1}, \lambda) \cdot p(\lambda \mid \mathcal{X}^{n-1})}{p(\mathbf{X}_n \mid \mathcal{X}^{n-1})} \\ &= \frac{p(\mathbf{X}_n \mid \lambda) \cdot p(\lambda \mid \mathcal{X}^{n-1})}{\int_\Omega p(\mathbf{X}_n \mid \lambda) \cdot p(\lambda \mid \mathcal{X}^{n-1}) d\lambda}. \end{aligned} \tag{4}$$

Starting with the calculation of the posterior pdf from $p(\lambda \mid \mathcal{X}^0) = p(\lambda)$, a repeated use of (4) produces the sequence of densities $p(\lambda \mid \mathcal{X}^1)$, $p(\lambda \mid \mathcal{X}^2)$, and so forth. This provides a basis of making recursive Bayesian inference of parameters λ [35].

Unfortunately, the implementation of this learning procedure for incremental CDHMM training raises some serious computational difficulties because of the nature of the missing-data problem caused by the underlying hidden processes, i.e., the state mixture component label sequence and the state sequence of the Markov chain for an HMM. It is well known that there exist no reproducing (natural conjugate) densities [35], [8], [12] for CDHMM. To illustrate this problem more clearly, let us begin with $p(\lambda \mid \mathcal{X}^0) = p(\lambda)$ and consider what happens after a training utterance (sample) $\mathbf{X}$ is observed. For an observation sequence $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T)$, let $\mathbf{s} = (s_1, s_2, \cdots, s_T)$ be the unobserved state sequence, and $\mathbf{l} = (l_1, l_2, \cdots, l_T)$ be the associated sequence of the unobserved mixture component labels. The posterior pdf of λ after observing $\mathbf{X}$ is

$$\begin{aligned} p(\lambda \mid \mathbf{X}) \propto \sum_{\mathbf{s}} \sum_{\mathbf{l}} & \left\{ \pi_{s_1} \omega_{s_1 l_1} \mathcal{N}(\mathbf{x}_1 \mid m_{s_1 l_1}, r_{s_1 l_1}) \right. \\ & \left. \times \prod_{t=2}^{T} a_{s_{t-1} s_t} \omega_{s_t l_t} \mathcal{N}(\mathbf{x}_t \mid m_{s_t l_t}, r_{s_t l_t}) \right\} \cdot p(\lambda) \end{aligned} \tag{5}$$

where the summations are taken over all possible state and mixture component label sequences. So the exact posterior pdf $p(\lambda \mid \mathbf{X})$ is a weighted sum of the prior pdf $p(\lambda)$ which includes $(N \cdot K)^T$ terms. Successive computation of (4) introduces an ever-expanding combination of the previous posterior pdf's and thus quickly leads to the combinatorial explosion of terms. As a result, formal recursive Bayes learning procedures of this kind have been regarded as of purely academic interest. In order to make it more practical, some approximations are needed to alleviate the computational difficulties. The procedure proposed here is to apply the Bayes recursion of (4) incrementally, with one or more observation samples considered at a time. It is followed by a suitable approximation to the resulting posterior pdf so as to obtain recursive estimates of the hyperparameters of the approximate posterior pdf. This is typically accomplished by restricting the approximated pdf to be in the class of conjugate pdf's of the complete-data distributions.

III. Approximate Solution: Quasi-Bayes Learning

A. General Formulation

The Bayesian algorithm for learning about λ considered in this paper involves the specification of an initial *a priori* density for λ, and the subsequent recursive computation of the approximate posterior density. For the general case of CDHMM, in which both the mean and precision parameters are assumed to be random, the initial prior pdf of λ is assumed to be [12]

$$g(\lambda) = g(\lambda') \cdot \prod_{i=1}^{N} \prod_{k=1}^{K} g(m_{ik}, r_{ik}) \tag{6}$$

where

$$\begin{aligned} g(\lambda') \propto & \prod_{i=1}^{N} \\ & \left\{ [\pi_i]^{\eta_i - 1} \cdot \left(\prod_{j=1}^{N} [a_{ij}]^{\eta_{ij} - 1} \right) \cdot \left(\prod_{k=1}^{K} [\omega_{ik}]^{\nu_{ik} - 1} \right) \right\} \end{aligned} \tag{7}$$

is the product of a series of Dirichlet pdf (sometimes called multivariate beta pdf), and thus takes the special form of a matrix beta pdf [27] with sets of positive hyperparameters of $\{\eta_i\}$, $\{\eta_{ij}\}$, $\{\nu_{ik}\}$. If the Gaussian mixand has a full precision matrix, then $g(m_{ik}, r_{ik})$ is assumed to be a normal-Wishart density of the form [8]

$$\begin{aligned} g(m_{ik}, r_{ik}) \propto & |r_{ik}|^{(\alpha_{ik} - D)/2} \\ & \times \exp\left[-\frac{\tau_{ik}}{2} (m_{ik} - \mu_{ik})^t r_{ik} (m_{ik} - \mu_{ik}) \right] \\ & \times \exp\left[-\frac{1}{2} \mathrm{tr}(u_{ik} r_{ik}) \right] \end{aligned} \tag{8}$$

where $\{\tau_{ik}, \mu_{ik}, \alpha_{ik}, u_{ik}\}$ are the hyperparameters of the prior density such that $\alpha_{ik} > D - 1$, $\tau_{ik} > 0$, μ_{ik} is a vector of dimension D and u_{ik} is a $D \times D$ positive definite matrix. Here, $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. This class of prior distributions actually constitutes a conjugate family of the complete-data density and is denoted as $\mathcal{P}$. The following discussion and formulation will be based on the general assumption of full precision matrix case. However, many practical CDHMM-based speech recognition systems usually adopt the diagonal precision matrices. For completeness, we will also summarize the related formulation in Appendix A.

We propose in this paper, at each step of the recursive Bayes learning discussed in previous section, to approximate the true posterior distribution $p(\lambda \mid \mathcal{X}^n)$ by the "closest" tractable distribution $g(\lambda \mid \varphi^{(n)})$ within the given class $\mathcal{P}$,

where $\varphi^{(n)}$ denotes the updated hyperparameters after observing the sample $\mathbf{X}_n$. The approximate MAP estimation of CDHMM parameters at this time is then obtained by

$$\lambda^{(n)} = \arg\max_{\lambda} g(\lambda \mid \varphi^{(n)}). \tag{9}$$

The term "closest" here depends, of course, on the particular criterion adopted in making the approximation. From the viewpoint of density approximation, minimizing the Kullback–Leibler directed divergence of the approximate pdf from the exact posterior pdf will give an attractive solution

$$\varphi^{(n)} = \arg\min_{\varphi} \int p(\lambda \mid \mathcal{X}^n) \log \frac{p(\lambda \mid \mathcal{X}^n)}{g(\lambda \mid \varphi)} d\lambda, \quad g(\cdot \mid \varphi) \in \mathcal{P}. \tag{10}$$

This procedure has an interesting decision-theoretical justification, as that which minimizes the expected loss when the decision space consists of all available approximations and the utility function is a proper, local scoring rule [5]. Unfortunately, no explicit closed-form solution exists for this problem and a general optimization procedure is needed to get the hyperparameters estimate. Interested readers are referred to [6] for an example of such a Bayesian analysis of a simple mixture problem. Instead of direct use of above approximation procedure, we suggest and highlight here a method called quasi-Bayes (QB) learning, which is both conceptually simple and computationally effective.

B. Quasi-Bayes Learning

The quasi-Bayes procedure is an approximate solution that is motivated by aiming at achieving computational simplicity while still maintaining the flavor of the formal Bayes procedure. In the context of finite mixture distribution identification, the quasi-Bayes approach was originally proposed by Makov and Smith [25], [34] to conduct recursive Bayes estimation of the mixture coefficients while the mixture components are assumed fixed. In the sense that the approximate posterior distribution with a mean identical to that of the true posterior distribution, the convergence properties were established. We previously adopted this approach to on-line adaptation of the mixture coefficients in the SCHMM case [14], [15]. In the following, we will expand this method to the CDHMM case.

At each step of recursive Bayes learning, the proposed quasi-Bayes procedure approximates the resulting posterior distribution $p(\lambda \mid \mathcal{X}^n)$, by the "closest" tractable distribution $g(\lambda \mid \varphi^{(n)})$ within the given class $\mathcal{P}$, under the criterion that both distributions have the same mode. This idea is schematically illustrated in Fig. 2. More specifically, consider at time instant n, we have a training utterance $\mathbf{X}_n = (\mathbf{x}_1^{(n)}, \mathbf{x}_2^{(n)}, \cdots, \mathbf{x}_{T_n}^{(n)})$, and our prior knowledge about λ is approximated by $g(\lambda \mid \varphi^{(n-1)})$. Let $\mathbf{Y}_n = (\mathbf{X}_n, \mathbf{Z}_n)$ denote the associated complete-data and $\mathbf{Z}_n = (\mathbf{s}_n, \mathbf{l}_n)$ be corresponding missing data with $\mathbf{s}_n = (s_1^{(n)}, s_2^{(n)}, \cdots, s_{T_n}^{(n)})$ being the unobserved state sequence and $\mathbf{l}_n = (l_1^{(n)}, l_2^{(n)}, \cdots, l_{T_n}^{(n)})$ being the associated sequence of the unobserved mixture component labels. We get the *approximate MAP* estimate $\lambda^{(n)}$ of λ by repeating the following steps.

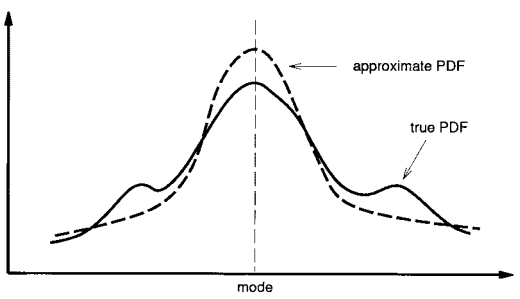

Fig. 2. Schematic illustration of quasi-Bayes procedure: The true posterior distribution is approximated by a simpler distribution under the criterion that both distributions have the same mode.

E-step: Compute

$$R\big(\lambda \mid \lambda^{(n-1+\frac{m-1}{M})}\big) = \rho \cdot \log g(\lambda \mid \varphi^{(n-1)}) + E\big[\log p(\mathbf{Y}_n \mid \lambda) \mid \mathbf{X}_n, \lambda^{\left(n-1+\frac{m-1}{M}\right)}\big] \tag{11}$$

where $0 < \rho \leq 1$ is a forgetting factor and $\rho = 1$ means that there is no forgetting.

M-step: Choose

$$\lambda^{(n-1+m/M)} = \arg\max_{\lambda} R\big(\lambda \mid \lambda^{(n-1+\frac{m-1}{M})}\big) \tag{12}$$

where $m = 1, 2, \cdots, M$ is the iteration index and M is the total iterations performed.

If the initial prior knowledge is too strong or after a lot of adaptation data have been incrementally processed, the new adaptation data usually have only a small impact on parameters updating in incremental training. To continuously track the variations of the model parameters corresponding to the new data, some *forgetting mechanism* is needed to reduce the effect of past observations relative to the new input data. Here we propose an *exponential forgetting* scheme by using a forgetting coefficient ρ as shown in (11). This is analogous to that proposed in [40] and [18].

By choosing the initial prior pdf to be the conjugate family of the *complete-data* density, it can be similarly verified as in [12] that with an appropriate normalization factor C, $C \cdot \exp\{R(\lambda \mid \lambda^{(n-1+\frac{m-1}{M})})\}$ belongs to the same distribution family as $g(\cdot)$, thus is denoted as $g(\lambda \mid \hat{\varphi})$ with the hyperparameters $\hat{\varphi}$ detailed as the following:

$$\hat{\eta}_i = \rho \cdot \big(\eta_i^{(n-1)} - 1\big) + 1 + \gamma_1(i) \tag{13}$$

$$\hat{\eta}_{ij} = \rho \cdot \big(\eta_{ij}^{(n-1)} - 1\big) + 1 + \sum_{t=1}^{T} \gamma_t(i,j) \tag{14}$$

$$\hat{\nu}_{ik} = \rho \cdot \big(\nu_{ik}^{(n-1)} - 1\big) + 1 + c_{ik} \tag{15}$$

$$\hat{\tau}_{ik} = \rho \tau_{ik}^{(n-1)} + c_{ik} \tag{16}$$

$$\hat{\mu}_{ik} = \frac{\rho \tau_{ik}^{(n-1)} \mu_{ik}^{(n-1)} + c_{ik} \bar{\mathbf{x}}_{ik}}{\rho \tau_{ik}^{(n-1)} + c_{ik}} \tag{17}$$

$$\hat{\alpha}_{ik} = \rho \cdot \big(\alpha_{ik}^{(n-1)} - D\big) + D + c_{ik} \tag{18}$$

$$\hat{u}_{ik} = \rho u_{ik}^{(n-1)} + S_{ik} + \frac{\rho \tau_{ik}^{(n-1)} c_{ik}}{\rho \tau_{ik}^{(n-1)} + c_{ik}} \cdot \big(\bar{\mathbf{x}}_{ik} - \mu_{ik}^{(n-1)}\big)\big(\bar{\mathbf{x}}_{ik} - \mu_{ik}^{(n-1)}\big)^t \tag{19}$$

where

$$\gamma_t(i,j) = \Pr(s_t = i, s_{t+1} = j \,|\, \mathbf{X}, \lambda) \quad 1 \le t \le T-1 \tag{20}$$

$$\gamma_t(i) = \Pr(s_t = i \,|\, \mathbf{X}, \lambda) \quad 1 \le t \le T \tag{21}$$

$$\zeta_t(i,k) = \Pr(s_t = i, l_t = k \,|\, \mathbf{X}, \lambda) \quad 1 \le t \le T \tag{22}$$

$$c_{ik} = \sum_{t=1}^{T} \zeta_t(i,k) \tag{23}$$

$$\bar{\mathbf{x}}_{ik} = \sum_{t=1}^{T} \zeta_t(i,k)\mathbf{x}_t / c_{ik} \tag{24}$$

$$S_{ik} = \sum_{t=1}^{T} \zeta_t(i,k)(\mathbf{x}_t - \bar{\mathbf{x}}_{ik})(\mathbf{x}_t - \bar{\mathbf{x}}_{ik})^t \tag{25}$$

and these terms can be computed efficiently by the forward-backward algorithm [17], [29]. Note that for notational simplicity, we have dropped the related subscripts and/or superscripts which indicate the iteration index and training sample index. The EM reestimation formulas of the CDHMM parameters can thus be derived by taking the mode of $g(\lambda \,|\, \hat{\varphi})$ and are shown as follows:

$$\hat{\pi}_i = \frac{\hat{\eta}_i - 1}{\sum_{j=1}^{N}(\hat{\eta}_j - 1)} = \frac{\rho \cdot \left(\eta_i^{(n-1)} - 1\right) + \gamma_1(i)}{\sum_{j=1}^{N}\left[\rho \cdot \left(\eta_j^{(n-1)} - 1\right) + \gamma_1(j)\right]} \tag{26}$$

$$\hat{a}_{ij} = \frac{\hat{\eta}_{ij} - 1}{\sum_{k=1}^{N}(\hat{\eta}_{ik} - 1)} = \frac{\rho \cdot \left(\eta_{ij}^{(n-1)} - 1\right) + \sum_{t=1}^{T}\gamma_t(i,j)}{\sum_{k=1}^{N}\left[\rho \cdot \left(\eta_{ik}^{(n-1)} - 1\right) + \sum_{t=1}^{T}\gamma_t(i,k)\right]} \tag{27}$$

$$\hat{\omega}_{ik} = \frac{\hat{\nu}_{ik} - 1}{\sum_{j=1}^{K}(\hat{\nu}_{ij} - 1)} = \frac{\rho \cdot \left(\nu_{ik}^{(n-1)} - 1\right) + \sum_{t=1}^{T}\zeta_t(i,k)}{\sum_{j=1}^{K}\left[\rho \cdot \left(\nu_{ij}^{(n-1)} - 1\right) + \sum_{t=1}^{T}\zeta_t(i,j)\right]} \tag{28}$$

$$\hat{m}_{ik} = \hat{\mu}_{ik} \tag{29}$$

$$\hat{r}_{ik}^{-1} = (\hat{\alpha}_{ik} - D)^{-1} \cdot \hat{u}_{ik} = \frac{\rho u_{ik}^{(n-1)} + \rho\tau_{ik}^{(n-1)}\left(\hat{m}_{ik} - \mu_{ik}^{(n-1)}\right)\left(\hat{m}_{ik} - \mu_{ik}^{(n-1)}\right)^t}{\rho\left(\alpha_{ik}^{(n-1)} - D\right) + \sum_{t=1}^{T}\zeta_t(i,k)} + \frac{\sum_{t=1}^{T}\zeta_t(i,k)(\mathbf{x}_t - \hat{m}_{ik})(\mathbf{x}_t - \hat{m}_{ik})^t}{\rho\left(\alpha_{ik}^{(n-1)} - D\right) + \sum_{t=1}^{T}\zeta_t(i,k)}. \tag{30}$$

By repeating the above EM iteration, we can get a series of approximate pdf $g(\lambda \,|\, \hat{\varphi})$ whose mode is approaching to the mode[1] of the true posterior pdf

$$p(\lambda \,|\, \mathbf{X}_n) = \frac{p(\mathbf{X}_n \,|\, \lambda) \cdot g(\lambda \,|\, \varphi^{(n-1)})}{\int_\Omega p(\mathbf{X}_n \,|\, \lambda) \cdot g(\lambda \,|\, \varphi^{(n-1)}) d\lambda}. \tag{31}$$

Thus, the hyperparameters $\varphi^{(n)}$ are obtained at the last (actually Mth) EM iteration by using (13)–(19) to satisfy

$$g(\lambda \,|\, \varphi^{(n)}) \propto \exp\left\{R\left(\lambda \,|\, \lambda^{(n-1+\frac{M-1}{M})}\right)\right\} \tag{32}$$

and the CDHMM parameters $\lambda^{(n)}$ are updated accordingly.

[1] Strictly speaking, the EM algorithm can only guarantee the mode of the approximate pdf to approach a local maximum of the above true posterior pdf.

C. Discussion

The above forward-backward type procedure can be easily extended to a segmental (or Viterbi) one by replacing (20)–(22) with

$$\gamma_t(i,j) = \delta(s_t - i)\delta(s_{t+1} - j) \tag{33}$$

$$\gamma_t(i) = \delta(s_t - i), \tag{34}$$

$$\zeta_t(i,k) = \gamma_t(i) \cdot \frac{\omega_{ik}\mathcal{N}(\mathbf{x}_t \,|\, m_{ik}, r_{ik})}{\sum_{j=1}^{K}\omega_{ij}\mathcal{N}(\mathbf{x}_t \,|\, m_{ij}, r_{ij})} \tag{35}$$

where $\mathbf{s} = (s_1, s_2, \cdots, s_T)$ is the most likely state sequence corresponding to observation sequence $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T)$, and $\delta(\cdot)$ denotes the Kronecker delta function.

In the above quasi-Bayes learning framework, if each time obtaining a training sample, only one EM iteration is performed and no forgetting is activated to update the CDHMM parameters and the associated hyperparameters, then the whole incremental quasi-Bayes learning process becomes the following recursive EM version for approximate MAP estimate originally suggested by Titterington in [37] as follows.

E-Step: Compute

$$L_n(\lambda) = E[\log p(\mathbf{Y}_n \mid \lambda) \mid \mathbf{X}_n, \lambda^{(n-1)}] + L_{n-1}(\lambda). \tag{36}$$

M-Step: Choose $\lambda = \lambda^{(n)}$ to maximize $L_n(\lambda)$ and also update hyperparameters to get $\varphi^{(n)}$.

In (36), one can initialize using

$$L_0(\lambda) = \log p(\lambda) = \log g(\lambda \,|\, \varphi^{(0)}) \tag{37}$$

where $p(\cdot)$ is the initial prior density for λ, with mode $\lambda^{(0)}$ and hyperparameters $\varphi^{(0)}$. This also shows that the quasi-Bayes procedure proposed in this paper is truly both computationally efficient and retains the flavor of the formal Bayes solution.

We have discussed the incremental training procedure to process one utterance at a time. Actually, the quasi-Bayes learning framework is flexible enough to include the batch or block mode learning as a special case. If the application permitted, one can also update the parameters by taking observations in batches, small enough to ensure that the computational requirements of the related Bayesian updating is within reasonable limits, so that the user will not be aware of the long delay. To an extreme, one can update the parameters by using all of the history data and the initial prior distribution. In this case, the QB method will degenerate to be the conventional batch mode MAP estimate.

The QB framework can also be used to implement an incremental version of the ML estimate of CDHMM. Given all of the training data, one first runs one batch-mode EM (Baum–Welch) iteration, and thus gets an initial prior pdf estimate by using (38)–(44) in the next section. Starting from this initial prior pdf, one can go through the training data again by using QB framework to incrementally update the related parameters. After the pass of the whole training data (called one *epoch*), one can *refresh* the posterior pdf by using (45)–(51) in the next section and then *feedback* the refreshed pdf to be the initial prior pdf (called *prior/posterior feedback*). The whole process can be repeated until convergence. In this regard, we wish to draw the reader's attention to the concurrent

and independent work of Gotoh *et al.* [13], who have used a similar method as the above quasi-Bayes learning procedure from a different viewpoint of speeding up the convergence of CDHMM training by using the above incremental algorithm instead of standard batch training one. In their work, however, they did not emphasize the underlying approximate nature of the updated posterior distribution to the exact one, thus failed to provide a sound formulation of the forgetting mechanism, albeit their awareness of its importance. We think this insight is important for developing other alternative methods as well as further studying some important issues such as the asymptotic convergence properties and the associated regularity conditions that have yet to be resolved. The existing literature on this topic, together with the ideas presented in this paper, should provide a starting point for such analyses. In fact, based on the general approximation theory discussed in Section III-A, apart from the QB learning method, we have also developed some other inference procedures which can be viewed, in a unified manner, as approximations to the formal recursive Bayesian solution demonstrated in Section II. We will report those results elsewhere. In the following sections, we will show by a series of experiments that the proposed QB algorithm does converge to a reasonable solution in terms of improving speech recognition rate. Before that, in next section, some important implementation issues will be first discussed.

IV. Implementation Issues

A. Initial Hyperparameter Estimation

In previous sections, the initial prior density $g(\lambda \mid \varphi^{(0)})$ is assumed to be a member of a preassigned family of prior distributions. In a strict Bayesian approach, the hyperparameter vector $\varphi^{(0)}$ of this family of pdf's $\{g(\cdot \mid \varphi^{(0)})\}$ is also assumed known based on a subjective knowledge about λ. In reality, it is difficult to possess a complete knowledge of the prior distribution. One solution is to adopt the *empirical Bayes* (EB) approach [32], [26] to estimate the initial hyperparameters $\varphi^{(0)}$.

Prior density estimation and the choice of density parameters depend on the particular application of interest. In the speaker adaptation (SA) application presented later in this study, the initial prior density $g(\lambda \mid \varphi^{(0)})$ represents the initial information of the variability of a certain model among a set of different speakers. Taking the empirical Bayes approach, the speaker-independent (SI) training data set $\mathcal{X}^{(\mathrm{SI})}$ for estimating hyperparameters $\varphi^{(0)}$ can be divided into different subsets $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_Q$ correspond to Q different speakers or speaker groups so that each token of the SI training data is associated with a speaker (group) ID. With those clustered training data, one can estimate Q sets of HMM's $\tilde{\Lambda} = (\tilde{\lambda}_1, \tilde{\lambda}_2, \cdots, \tilde{\lambda}_Q)$ with the classical Baum–Welch or segmental k-means algorithm. One can also perform an SI training at first by using all of the training data. At the last iteration of SI training, with the help of the speaker (group) ID information associated with each training token, one can accumulate Q sets of related statistics and thus correspondingly derive Q sets of HMM's. One then pretends to view $\{\tilde{\lambda}_i\}$ as a collection of random observations from the density $g(\lambda \mid \varphi^{(0)})$. When enough SI training data are available, the method of moment as discussed in [14] can be used to estimate the hyperparameters $\varphi^{(0)}$. Otherwise, the following *ad hoc* method can also be used.

By viewing the last SI training iteration as MAP estimation with the *noninformative prior*, we get the estimate of the hyperparameters $\varphi^{(0)}$ in the same spirit as in quasi-Bayes learning framework as follows:

$$\eta_i^{(0)} = 1 + \epsilon_1 \cdot \gamma_1^{(\mathrm{SI})}(i) \tag{38}$$

$$\eta_{ij}^{(0)} = 1 + \epsilon_1 \cdot \sum_t \gamma_t^{(\mathrm{SI})}(i,j) \tag{39}$$

$$\nu_{ik}^{(0)} = 1 + \epsilon_1 \cdot \sum_t \zeta_t^{(\mathrm{SI})}(i,k) \tag{40}$$

$$\tau_{ik}^{(0)} = \epsilon_1 \cdot \sum_t \zeta_t^{(\mathrm{SI})}(i,k) \tag{41}$$

$$\mu_{ik}^{(0)} = m_{ik}^{(\mathrm{SI})} \tag{42}$$

$$\alpha_{ik}^{(0)} = D + \epsilon_1 \cdot \sum_t \zeta_t^{(\mathrm{SI})}(i,k) \tag{43}$$

$$u_{ik}^{(0)} = \epsilon_1 \cdot \left(r_{ik}^{(\mathrm{SI})}\right)^{-1} \sum_t \zeta_t^{(\mathrm{SI})}(i,k) \tag{44}$$

where $0 < \epsilon_1 \leq 1$ is a weighting coefficient to control the importance of the prior knowledge or to balance the contribution between the SI training data and the adaptation data. This parameter can be specified by a user. It can also be determined *a posteriori* by measuring in some way the similarity between the coming adaptation data and the existing models. This could be a topic for further research. By choosing these estimators for the hyperparameters, we sacrifice the ability of the prior density to accurately model the interspeaker variability but we obtain more robust estimators in case when only insufficient SI training data are available. On the other hand, if a sufficient amount of SI training data is available, the first method (method of moment) can lead to a more accurate hyperparameter estimate by considering the interspeaker statistics. In this way, the importance of the prior knowledge is determined directly by the available SI training data.

B. Refreshing Hyperparameters

In the Bayesian adaptive learning framework, the adaptation effects depend heavily on the suitability of the prior distribution to the new data. The motivation of on-line adaptation (OLA) is to adapt a recognizer continuously to the coming block of new utterances, and then hope this adapted recognizer will do better for the next test utterance than the one without applying OLA. We usually start OLA from a general initial model (e.g., SI model), and then continuously adapt to the new data. As discussed in Section III, if the initial prior knowledge is too strong, or after a lot of adaptation data have been incrementally processed, some forgetting mechanism is needed to help continuously track the variations of the model parameters corresponding to the new data. There are many ways to implement the forgetting mechanism to reduce the mismatch between updated posterior distribution and the coming data, and to track the varying conditions. The

exponential forgetting is expected to be helpful for handling the slow changes of acoustic conditions between consecutive utterances by deemphasizing the contribution of the history data. If at a certain time the condition changes abruptly, say, a change of speaker, then the prior (or the updated posterior) distribution may not provide much useful information for this new speaker and thus deteriorate the efficacy of the OLA. In this case, exponential forgetting may be too slow and not able to handle such fast changes. Refreshing the hyperparameters may be more helpful for *fast forgetting*. The simplest way is to back-off to the general (e.g., SI) initial models, which usually provide a reasonable performance and a robust initial hyperparameters' estimate. If the situation permitted, it will be helpful to maintain multiple sets of prior (updated posterior) distributions and select upon some criterion the best one to refresh. Finally, we can also normalize the updated hyperparameters themselves to deemphasize their contributions to the new adaptation data as follows:

$$\hat{\eta}_i = 1 + \epsilon_2 \cdot (\eta_i - 1) \tag{45}$$

$$\hat{\eta}_{ij} = 1 + \epsilon_2 \cdot (\eta_{ij} - 1) \tag{46}$$

$$\hat{\nu}_{ik} = 1 + \epsilon_2 \cdot (\nu_{ik} - 1) \tag{47}$$

$$\hat{\tau}_{ik} = \epsilon_2 \cdot \tau_{ik} \tag{48}$$

$$\hat{\mu}_{ik} = \mu_{ik} \tag{49}$$

$$\hat{\alpha}_{ik} = D + \epsilon_2 \cdot (\alpha_{ik} - D) \tag{50}$$

$$\hat{u}_{ik} = \epsilon_2 \cdot u_{ik} \tag{51}$$

where $0 < \epsilon_2 \leq 1$ is a weighting coefficient to control the *degree* of the forgetting.

V. Speaker Adaptation Experiments

A. Experimental Setup

To examine the viability of the proposed techniques, the incremental quasi-Bayes adaptive learning framework is applied to on-line speaker adaptation. We report on a series of recognition experiments using a vocabulary of the 26-letter English alphabet. Two severely mismatched speech databases were used for evaluating the adaptation algorithm. These two corpora, the OGI ISOLET and the TI46, were recorded at two separate sites with a time gap of ten years. The speech data were digitized at sampling rates of 16 kHz with 16-b quantization and 12.5 kHz with 12-b quantization respectively. The ISOLET corpus was recorded with a Sennheiser HMD 224 close-talking noise-cancelling microphone and the TI46 corpus was recorded with an Electro-Voice RE-16 cardoid dynamic microphone positioned two inches from the speaker's mouth. They have, therefore, very different acoustic characteristics. The speech data in the two corpora are lowpass-filtered at 3.3 kHz and downsampled to 8 kHz so that hopefully, they will become more compatible to each other. For speaker independent training and initial prior density estimation, the OGI ISOLET database was used. It consists of 150 speakers, 75 females and 75 males, each speaking each of the letters twice. For incremental speaker adaptive training and testing, the English alphabet subset of the TI46 isolated word corpus was used. It was produced by 16 speakers, eight females, and eight males. Among them, data from four males were incomplete. Therefore only 12 speakers were used in this study. Each person uttered each of the letters 26 times. Ten of them were collected in the same session. They are collectively denoted as DAT1 in this study. The remaining 16 tokens denoted as DAT2 were collected in eight different sessions in which two tokens of each letter were collected in each session. We divided DAT2 equally into two sets denoted, respectively, as DAT4 and DAT5.

For all the experiments, each letter in the vocabulary was modeled by a single left-to-right five-state CDHMM with arbitrary state skipping. Each state had four Gaussian mixture components with each component having a diagonal covariance matrix. Each feature vector used in this study consisted of 12 bandpass-liftered LPC-derived cepstral coefficients with a 30-ms frame length and a 10-ms frame shift [20]. Although there are other alternatives (e.g., [33]), only utterance-based cepstral mean subtraction (CMS) was applied for acoustic normalization. In all of the experiments, three EM iterations were performed for batch mode MAP training and incremental QB training. The initial hyperparameters were estimated by using the second method discussed in Section IV-A. In the particular experiments here, the weighting coefficient ϵ_1 was chosen to be $1/W$ with W being the number of SI training tokens corresponding to each HMM. This was equivalent to control the importance of the initial prior knowledge to be comparable with the contribution from a single training token. In recognition, the decision rule determined the recognized letter as the one which attained the highest forward-backward probability.

In the following subsections, we study the convergence property of the algorithm, the effects of different initial conditions, and the utility of the forgetting mechanism. All of the experiments were performed in a supervised mode.

B. Convergence Property

To examine the convergence property of the proposed algorithm, we started with the SI initial models and performed supervised on-line adaptation by using DAT4 as the adaptation set. After each OLA step, we test the recognizer by using DAT5 as the testing set. We plot in Fig. 3 the OLA performance, averaged over 12 speakers, as a function of the number of adaptation tokens per letter (labeled as "si-ini-ol-dat4"), although OLA is actually performed after each utterance is available. For comparison, the adaptation results by using batch MAP training method are also plotted and labeled as "si-ini-map-dat4." The results showed that both OL and batch MAP adaptation can consistently and continuously improve the recognition performance when more and more adaptation data were available. The small performance difference between the two methods confirms that the proposed quasi-Bayes approximation to the true posterior distribution is viable and efficient. One advantage of the OL implementation over its batch counterpart lies in its computational efficiency and reduced storage requirements. More importantly, by incrementally updating hyperparameters and introducing the forgetting mechanism, the algorithm is truly adaptive in nature and can

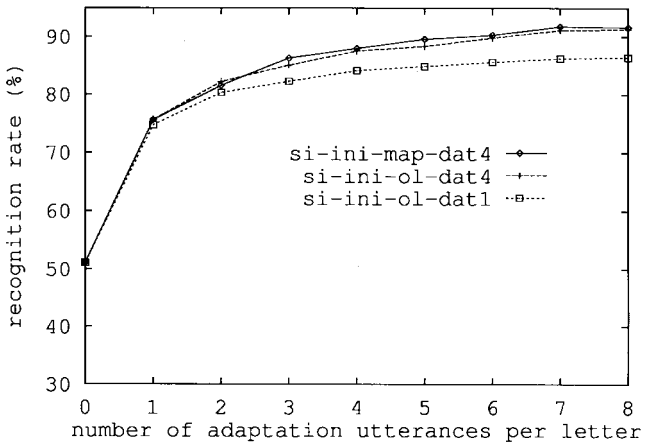

Fig. 3. Convergence and performance comparison of on-line and batch adaptation (starting from SI initial models, $\rho = 1$, averaged over 12 speakers).

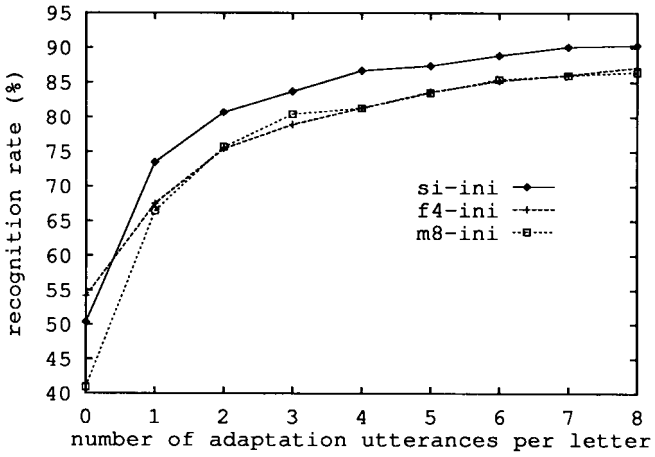

Fig. 4. Performance comparison under different initial conditions as a function of number of adaptation data ($\rho = 1$, averaged over seven female speakers).

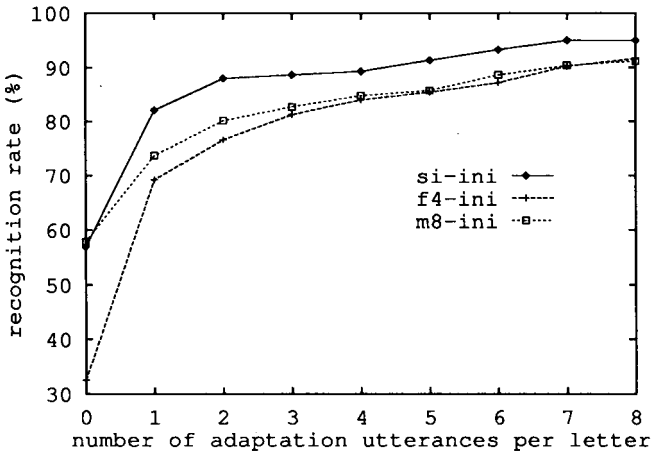

Fig. 5. Performance comparison under different initial conditions as a function of number of adaptation data ($\rho = 1$, averaged over three male speakers).

continuously track the changing conditions. We will provide more experimental evidences in the following subsections. Before that, we also plotted in Fig. 3 the OLA results by using DAT1 as adaptation set and testing on DAT5 to show the session effects between adaptation and testing data. The corresponding performance curve is labeled as "si-ini-ol-dat1." As expected, it is inferior to "si-ini-ol-dat4," because DAT1 and DAT5 were collected in completely different sessions. Whereas for each testing token in DAT5, there correspondingly exists an adaptation token in DAT4 coming from the same session.

C. Effects of Initial Conditions

In speaker adaptation based on Bayesian learning framework, one hopes to use prior distribution of CDHMM parameters to represent the information of the variability of a certain model among different speakers, so SA effects depend heavily on the suitability of the prior distribution to the new speaker. To show effects of OLA under different initial conditions, apart from starting OLA from SI model, other initial conditions are also tried. Specifically here, we first arbitrarily choose two speakers, one female (f4) and one male (m8). Starting from SI models, we perform OLA, respectively, on f4 and m8 by using DAT1 as adaptation data. Then start from these SA models, we perform OLA on other ten speakers by using DAT4 as adaptation data. Once again, DAT5 is used to test the recognizer after each OLA step. In Fig. 4, we plot the performance comparison averaged over seven female speakers under different initial conditions. We can see that the OLA performance from SA initials of f4 and m8 is inferior to the one from SI initial model. This is partly due to the severer mismatch between the prior distribution (e.g., for m8) and the new adaptation data, and partly because after OLA with DAT1 for f4 and m8, the updated hyperparameters represent too strong prior information in comparison with the contribution of new data from new speakers, especially when new adaptation data is insufficient. The latter is confirmed by the fact that when no new adaptation data is available, the recognition rate with SA initial models of f4 is better than the one with SI model, but the OLA performance starting from f4 initials is still inferior to the one from SI initials. This also confirms the necessity and importance of some kind of forgetting mechanism for the efficiency of OLA, especially in gender switching conditions. We will show in the next subsection that the introduction of this mechanism does help improve the OLA's efficacy. A similar observation can also be derived from the performance comparison averaged over three male speakers as shown in Fig. 5.

D. Effects of the Forgetting Mechanism

To examine the effect of the exponential forgetting factor in the case of slowly changing acoustic conditions, we performed for each speaker OLA starting from the SI initials with DAT1 as the adaptation data. Then we activated the exponential forgetting mechanism and continued OLA by using DAT4 as the adaptation data. After each OLA step we tested the recognizer with DAT5. The performance comparison by using different forgetting coefficients (fc) is plotted in Fig. 6. In the particular experimental setup here, the results showed that even with 18 tokens per letter, the performance has not saturated yet. Although the effect of the forgetting mechanism was

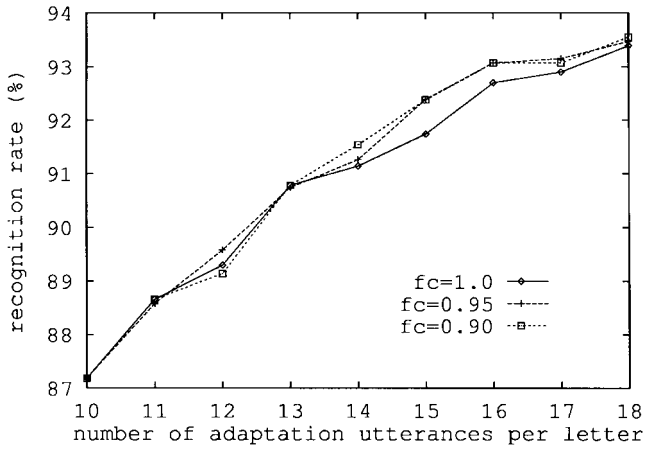

Fig. 6. Performance comparison with different forgetting coefficients to cope with slow varying conditions (averaged over 12 speakers).

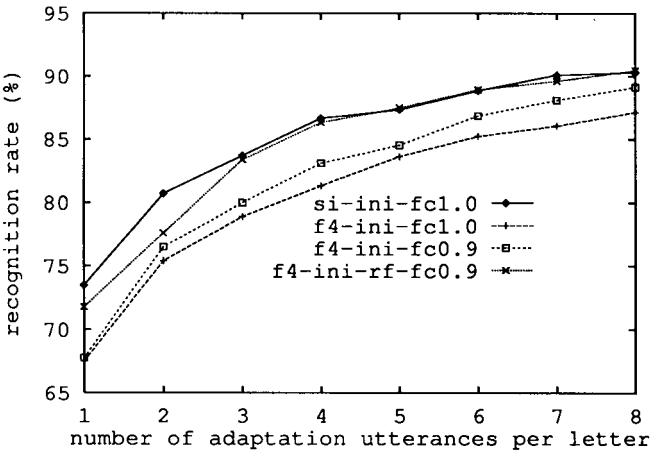

Fig. 7. Performance comparison with different forgetting schemes to cope-with fast varying conditions (starting from SA initials of f4, averaged over seven female speakers).

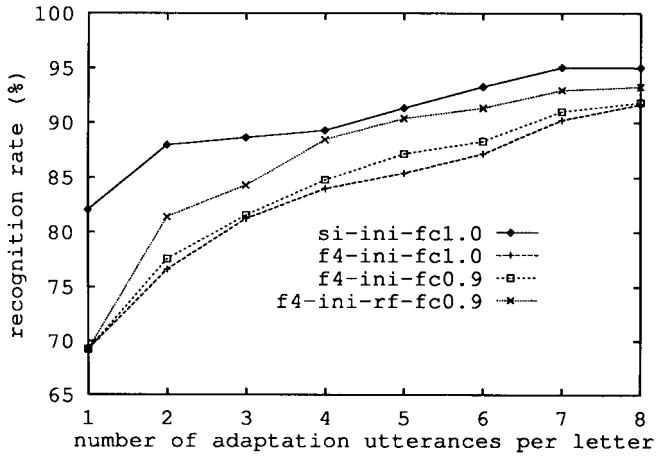

Fig. 8. Performance comparison with different forgetting schemes to cope-with fast varying conditions (starting from SA initials of f4, averaged over three male speakers).

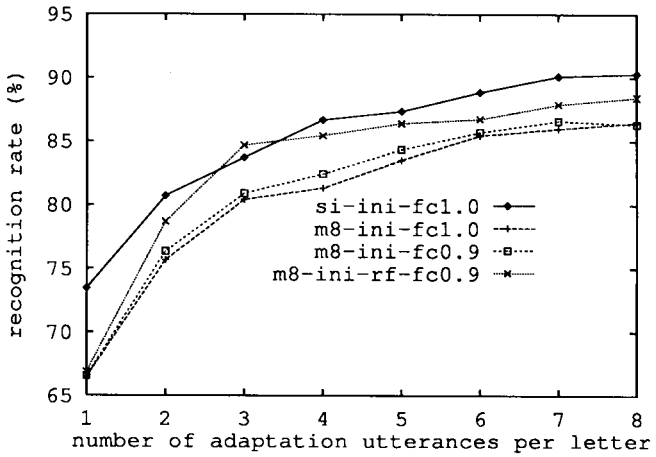

Fig. 9. Performance comparison with different forgetting schemes to cope-with fast varying conditions (starting from SA initials of m8, averaged over seven female speakers).

small, we can still see some improvement by activating the forgetting mechanism. The smaller the forgetting coefficient, the faster the forgetting process converged. However, a smaller forgetting coefficient also means a less weight assigned to the latest history data during adaptation. This may sometimes hamper the performance improvement, especially in the cases of insufficient and/or slowly changing adaptation data. The optimal value of forgetting coefficient should be situation dependent and its effect will be more apparent when large amount of adaptation data have been processed. Unfortunately, with the corpus we were using, we did not have enough data to conduct such a simulation.

To examine the effect of the forgetting mechanism in the case of abrupt switch of conditions (e.g., change of user), as an example, in Fig. 7, we plot the performance comparison averaged over seven female speakers to show the effects of different forgetting schemes. Starting from the SA initials of f4, when no forgetting mechanism, the OLA performance ("f4-ini-fc1.0") is much inferior to the one from the SI initials ("si-ini-fc1.0"). By activating exponential forgetting ("f4-ini-fc0.9"), we can see it helped improve the OLA performance, but it seemed not enough. By further including the mechanism of refreshing hyperparameters (the weighting coefficient ϵ_2 was chosen to be $1/W_1$ with W_1 being the number of SA tokens corresponding to each HMM used for speaker f4 adaptation), we can see that OLA performance ("f4-ini-rf-fc0.9") was improved significantly and quickly approach to the one obtained with the SI initials. Similar results were observed for the cases of starting from the SA initials of m8 as well as the performance comparison averaged over male speakers as shown in Figs. 8, 9, and 10. Note that all of the above experiments were performed in a supervised mode. However, in some applications, the recognition system has to be run in an unsupervised mode. In this case, how to automatically determine when to refresh the priors is an important research topic.

VI. Discussion and Conclusion

In this paper, we have presented a theoretical framework of QB learning of CDHMM with Gaussian mixture state observation densities based on a unified view of approximate recursive Bayesian inference. The implied algorithm can be

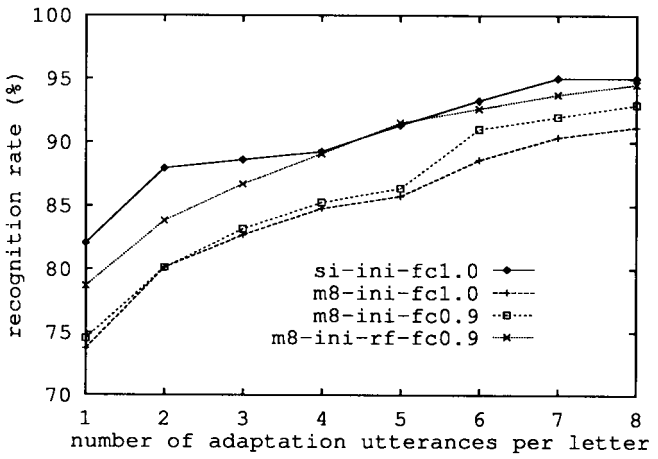

Fig. 10. Performance comparison with different forgetting schemes to cope-with fast varying conditions (starting from SA initials of m8, averaged over 3 male speakers).

adaptive in nature so that it can be used to perform a full-scale on-line adaptive learning of the CDHMM parameters only using the current available data to continuously track the varying acoustic conditions. To examine the viability of the proposed algorithm, the QB learning framework is applied to an on-line speaker adaptation application using the 26-letter English alphabet vocabulary. In a series of comparative experiments, we studied the convergence property of the algorithm, the effects of different initial conditions, and the utility of the forgetting mechanism. We have found the following.

- The QB learning algorithm does converge to a reasonable solution in terms of improving recognition rate and has a similar behavior with the batch MAP algorithm in cases when no forgetting mechanism is imposed.
- A good initial prior distribution is a key for improving the efficacy of on-line adaptation.
- The forgetting mechanism is useful in handling the slow changes of acoustic conditions between consecutive utterances and coping with the abrupt switch of speaking conditions (e.g., change of user). Two methods, exponential forgetting and hyperparameter refreshing, are proposed, and their usefulness has been confirmed.

In the experimental study of this paper, OLA is supervised, i.e., the true transcription of the adaptation data was assumed known. In practice, for on-line applications, unsupervised adaptation is often more realistic than the supervised one. The efficiency and effectiveness of unsupervised adaptation depend on the quality of the recognizer being used. If the current recognizer gives poor recognition results, then the supervision information is often wrong. This often results in an adapted model which gives worse performance than that obtained without adaptation. Moreover, OLA only uses the history data once. If it performs wrong adaptation at the very beginning, the system may diverge. In order to make OLA also work well in an unsupervised mode, it is desirable to minimize the effects of wrong supervision. Research along this line of thought is in progress. Another issue is about improving adaptation efficiency using data collected in mismatch acoustic conditions. In the acoustic normalization module in Fig. 1, we have only used the so-called blind equalization method in this study. Combining other acoustic normalization techniques with the current on-line adaptation framework is an important research topic. On the other hand, to improve the rate of adaptation when the data amount is insufficient, one may combine on-line Bayes adaptation with other methods such as vector field smoothing (VFS) technique [36], [38], transformation-based methods [10], [23], and other Bayesian techniques [19], where the dependency or correlation between HMM's is exploited to help adjust those HMM parameters without adaptation data. Actually, by combining with the so called extended MAP method [19], we have extended the current QB framework to cope with the correlated CDHMM's [16]. As a final remark, although the experiments discussed in this study are for speaker adaptation, the same formulation can also be used to handle varying channels, environments, and transducer mismatch problems in speech as well as speaker and other pattern recognition problems.

Appendix A
Formulation for Diagonal Precision Matrix Case

If the Gaussian mixand in (1) has a diagonal precision matrix, then $g(m_{ik}, r_{ik})$ is assumed to be a product of normal-gamma densities [8] with the form

$$g(m_{ik}, r_{ik}) \propto \prod_{d=1}^{D} r_{ikd}^{(\alpha_{ikd}-1/2)} \exp\left[-\frac{1}{2}\tau_{ikd} r_{ikd}(m_{ikd}-\mu_{ikd})^2\right] \times \exp[-\beta_{ikd} r_{ikd}] \tag{52}$$

where the hyperparameters τ_{ikd}, α_{ikd}, $\beta_{ikd} > 0$, $d = 1, 2, \cdots, D$. The updating formulas of hyperparameters $\{\eta_i\}$, $\{\eta_{ij}\}$, $\{\nu_{ik}\}$ have the same form as (13)–(15) and the remaining ones are as follows:

$$\hat{\tau}_{ikd} = \rho\tau_{ikd}^{(n-1)} + c_{ik} \tag{53}$$

$$\hat{\mu}_{ikd} = \frac{\rho\tau_{ikd}^{(n-1)}\mu_{ikd}^{(n-1)} + c_{ik}\bar{x}_{ikd}}{\rho\tau_{ikd}^{(n-1)} + c_{ik}} \tag{54}$$

$$\hat{\alpha}_{ikd} = \rho \cdot \left(\alpha_{ikd}^{(n-1)} - 0.5\right) + 0.5 + 0.5c_{ik} \tag{55}$$

$$\hat{m}_{ikd} = \hat{\mu}_{ikd} \tag{59}$$

$$\hat{r}_{ikd}^{-1} = \frac{2\rho\beta_{ikd}^{(n-1)} + \rho\tau_{ikd}^{(n-1)}\left(\hat{m}_{ikd} - \mu_{ikd}^{(n-1)}\right)^2 + \sum_{t=1}^{T}\zeta_t(i,k)(x_{td} - \hat{m}_{ikd})^2}{\rho\left(2\alpha_{ikd}^{(n-1)} - 1\right) + \sum_{t=1}^{T}\zeta_t(i,k)}. \tag{60}$$

$$\hat{\beta}_{ikd} = \rho\beta_{ikd}^{(n-1)} + \frac{1}{2}S_{ikd} + \frac{\rho\tau_{ikd}^{(n-1)}c_{ik}}{2\left(\rho\tau_{ikd}^{(n-1)} + c_{ik}\right)} \times \left(\bar{x}_{ikd} - \mu_{ikd}^{(n-1)}\right)^2 \quad (56)$$

where

$$\bar{x}_{ikd} = \sum_{t=1}^{T} \zeta_t(i,k)x_{td}/c_{ik} \quad (57)$$

$$S_{ikd} = \sum_{t=1}^{T} \zeta_t(i,k)(x_{td} - \bar{x}_{ikd})^2. \quad (58)$$

The updating formulas of CDHMM parameters $\{\pi_i\}$, $\{a_{ij}\}$, $\{\omega_{ik}\}$ have also the same form as (26)–(28), and the ones of $\{m_{ik}, r_{ik}\}$ are shown in (59) and (60), at the bottom of the previous page. Accordingly, the estimation formulas of the initial hyperparameters in (43) and (44) will be changed to

$$\alpha_{ikd}^{(0)} = 0.5 + 0.5\epsilon_1 \cdot \sum_t \zeta_t^{(\mathrm{SI})}(i,k) \quad (61)$$

$$\beta_{ikd}^{(0)} = 0.5\epsilon_1 \cdot \left(r_{ikd}^{(\mathrm{SI})}\right)^{-1} \sum_t \zeta_t^{(\mathrm{SI})}(i,k) \quad (62)$$

and the refreshing formulas of hyperparameters in (50) and (51) will become

$$\hat{\alpha}_{ikd} = 0.5 + \epsilon_2 \cdot (\alpha_{ikd} - 0.5) \quad (63)$$

$$\hat{\beta}_{ikd} = \epsilon_2 \cdot \beta_{ikd}. \quad (64)$$

ACKNOWLEDGMENT

The first author would like to thank Y. Yamazaki, President, ATR Interpreting Telecommunications Research Laboratories, and Y. Sagisaka, Head, Department 1 of the ATR-ITL, for their continuous support of this work.

REFERENCES

[1] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*. Norwell, MA: Kluwer, 1993.
[2] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1304–1312, 1974.
[3] L. E. Baum, "An inequality and associated maximization techniques in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, pp. 1–8, 1972.
[4] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic function of Markov chains," *Ann. Math. Stat.*, vol. 41, pp. 164–171, 1970.
[5] J. M. Bernardo, "Approximations in statistics from a decision-theoretical viewpoint," in *Probability and Bayesian Statistics*, R. Viertl, Ed. New York: Plenum, pp. 53–60, 1987.
[6] J. M. Bernardo and F. J. Giron, "A Bayesian analysis of simple mixture problems," *Bayesian Statistics 3*, J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Eds. Oxford, UK: Oxford Univ. Press, 1988, pp. 67–78.
[7] S. Cox and J. Bridle, "Unsupervised speaker adaptation by probabilistic spectrum fitting," in *Proc. ICASSP-89*, pp. 294–297.
[8] M. H. DeGroot, *Optimal Statistical Decisions*. New York: McGraw-Hill, 1970.
[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc, Ser. B*, vol. 39, no. 1, pp. 1–38.
[10] V. V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 5, pp. 357–366.
[11] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. ICASSP-96*, Atlanta, GA, May 1996, pp. 346–349.
[12] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 291–298, Apr. 1994.
[13] Y. Gotoh, M. M. Hochberg, D. J. Mashao, and H. F. Silverman, "Incremental MAP estimation of HMM's for efficient training and improved performance," in *Proc. ICASSP-95*, Detroit, MI, pp. I-457–I-460.
[14] Q. Huo, C. Chan, and C.-H. Lee, "Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 5, pp. 334–345.
[15] ———, "On-line adaptation of the SCHMM parameters based on the segmental quasi-Bayes learning for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 141–144, 1996.
[16] Q. Huo and C.-H. Lee, "On-line adaptive learning of the correlated continuous density hidden Markov models for speech recognition," in *Proc. ICSLP-96*, Philadelphia, PA, Oct. 1996, pp. 985–988.
[17] B.-H. Juang, S. E. Levinson, and M. M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of Markov chains," *IEEE Trans. Inform. Theory*, vol. IT-32, no. 2, pp. 307–309.
[18] V. Krishnamurthy and J. B. Moore, "On-line estimation of hidden Markov model parameters based on the Kullback-Leibler information measure," *IEEE Trans. Signal Processing*, vol. 41, no. 8, pp. 2557–2573.
[19] M. J. Lasry and R. M. Stern, "*A posteriori* estimation of correlated jointly Gaussian mean vectors," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, no. 4, pp. 530–535.
[20] C.-H. Lee, L. R. Rabiner, R. Pieraccini, and J. G. Wilpon, "Acoustic modeling for large vocabulary speech recognition," *Comput. Speech Lang.*, vol. 4, pp. 127–165, 1990.
[21] 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. Signal Processing*, vol. 39, pp. 806–814, Apr. 1991.
[22] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. ICASSP-96*, Atlanta, GA, pp. 353–356.
[23] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, pp. 171–185, 1995.
[24] L. R. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 729–734.
[25] U. E. Makov and A. F. M. Smith, "A quasi-Bayes unsupervised learning procedure for priors," *IEEE Trans. Inform. Theory*, vol. IT-23, no. 6, pp. 761–764.
[26] J. S. Maritz and T. Lwin, *Empirical Bayes Methods*, 2nd ed. London, UK: Chapman & Hall, 1989.
[27] J. J. Martin, *Bayesian Decision Problems and Markov Chains*. New York: Wiley, 1967.
[28] T. Matsuoka and C.-H. Lee, "A study of on-line Bayesian adaptation for HMM-based speech recognition," in *Proc. EUROSPEECH-93*, Berlin, Germany, pp. 815–818.
[29] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proc. IEEE*, vol. 77, no. 2, pp. 257–286.
[30] M. G. Rahim and B.-H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 19–30, Jan. 1996.
[31] M. G. Rahim, B.-H. Juang, W. Chou, and E. Buhrke, "Signal conditioning techniques for robust speech recognition," *IEEE Signal Processing Lett.*, vol. 3, pp. 107–109, Apr. 1996.
[32] H. Robbins, "The empirical Bayes approach to statistical decision problems," *Ann. Math. Stat.*, vol. 35, pp. 1–20, 1964.
[33] A. Sankar and C.-H. Lee, "A maximum likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, no. 3, pp. 190–202.
[34] A. F. M. Smith and U. E. Makov, "A quasi-Bayes sequential procedure for mixtures," *J. Roy. Stat. Soc., Ser. B*, vol. 40, no. 1, pp. 106–112.
[35] J. Spragins, "A note on the iterative application of Bayes' rule," *IEEE Trans. Inform. Theory*, vol. IT-11, no. 4, pp. 544–549.
[36] J. Takahashi and S. Sagayama, "Vector-field-smoothed Bayesian learning for incremental speaker adaptation," in *Proc. ICASSP-95*, Detroit, MI, pp. I-696–I-699.
[37] D. M. Titterington, "Recursive parameter estimation using incomplete data," *J. Roy. Stat. Soc., Ser. B*, vol. 46, no. 2, pp. 257–267.
[38] M. Tonomura, T. Kosaka, and S. Matsunaga, "Speaker adaptation based on transfer vector field smoothing using maximum *a posteriori* probability estimation," in *Proc. ICASSP-95*, Detroit, MI, pp. I-688–I-691.
[39] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in *Proc. ICASSP-96*, Atlanta, GA, pp. 339–341.

[40] E. Weinstein, M. Feder, and A. V. Oppenheim, "Sequential algorithms for parameter estimation based on the Kullback-Leibler information measure," *IEEE Trans. Acoust, Speech, Signal Processing*, vol. 38, no. 9, pp. 1652–1654.
[41] Y.-X. Zhao, "An acoustic-phonetic-based speaker adaptation technique for improving speaker-independent continuous speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 3, pp. 380–394.
[42] Y.-X. Zhao, "Self-learning speaker and channel adaptation based on spectral variation source decomposition," *Speech Commun.*, vol. 18, pp. 65–77, 1996.

Qiang Huo (M'95) received the B.Eng. degree from University of Science and Technology of China (USTC), China, in 1987, the M.Eng. degree from Zhejiang University, China, in 1989, and the Ph.D. degree from USTC in 1994, all in electrical engineering.

From 1986 to 1990, his research work focused on the hardware design and development for real-time digital signal processing, image processing and computer vision, and speech and speaker recognition. From 1991 to 1994, he was with the Department of Computer Science, University of Hong Kong, where he was involved in research on speech recognition. Since April 1995, he has been with ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan. His current major research interests include adaptive signal modeling and processing, speech recognition, speaker recognition, computational model for spoken dialogue processing, Chinese character recognition, and general pattern recognition theory.

Chin-Hui Lee (S'79-M'81-SM'90–F'97) received the B.S. degree from National Taiwan University, Taipei, in 1973, the M.S. degree from Yale University, New Haven, CT, in 1977, and the Ph.D. degree from University of Washington, Seattle, in 1981, all in electrical engineering.

In 1981, he joined Verbex Corporation, Bedford, MA, where he was involved in research on connected word recognition. In 1984, he became affiliated with Digital Sound Corporation, Santa Barbara, CA, where he engaged in research on speech coding, speech recognition, and signal processing for the development of the DSC-2000 Voice Server. Since 1986, he has been with AT&T Bell Laboratories, Murray Hill, NJ, where he is now a Distinguished Member of the Technical Staff and the Head of the Dialogue Systems Research Department at Bell Laboratories, Lucent Technologies. His current research interests include signal processing, speech modeling, adaptive and discriminative modeling, speech recognition, speaker recognition, and spoken dialogue processing. His research scope is reflected in a recently edited book *Automatic Speech and Speaker Recognition: Advanced Topics* (Norwell, MA: Kluwer, 1996).

Dr. Lee was he was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1991–1995) . He ws a member of the ARPA Spoken Language Processing Coordination Committee between 1991 and 1995. He has also served as a member of the Speech Technical Committee of the IEEE Signal Processing Society (SPS) since 1995. In 1996, he helped promote the newly formed SPS Multimedia Signal Processing Technical Committee (MMSP-TC) and is a member of the MMSP-TC). He is a recipient of the 1994 SPS Senior Award. He currently serves as the Chairman of the SPS Speech Technical Committee.