On-Line Adaptive Learning of the Correlated Continuous Density Hidden Markov Models for Speech Recognition

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

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

***Index Terms*— Automatic speech recognition, continuous density hidden Markov models, EM algorithm, recursive Bayesian estimation, speaker adaptation.**

I. Introduction

In the last decade, many advances have been made in the area of automatic speech recognition (ASR) (see e.g., [26] and other articles in [21]). However, it is also apparent that the performance of a speech recognizer often degrades drastically when acoustic mismatch between the testing and training conditions exists (see reviews, e.g., [9], [17], [22]). Most current recognition systems rely on a static design strategy in that all the knowledge sources needed in a system are acquired at the design phase and remain fixed during use. Since the design samples are often limited and the real conditions are always changing, this will inevitably result in some mismatch problems, and thus deteriorate the recognition performance. A better way is to acquire the knowledge dynamically. New information is constantly collected during development and use, and is incorporated into the system using adaptive learning algorithm.

Recently, Bayesian learning of hidden Markov model (HMM) parameters has been proposed and adopted in a number of adaptive speech recognition applications (e.g., [10], [12]–[14], [20]). A theoretical framework of Bayesian learning was first proposed by Lee *et al.* [20] 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., [10]) 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., [12]). It was shown that, for HMM-based speech recognition applications, the maximum a posteriori (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. This approach works in a batch adaptation mode using a history of all the adaptation data. A more attractive adaptation scheme is the so called on-line (or incremental, sequential) adaptation, which is able to update both the parameters of the prior and/or posterior distributions (called hyperparameters) and the HMM parameters themselves simultaneously upon the presentation of the latest adaptation data. 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. One such approach, called *quasi-Bayes* (QB) learning, was recently developed in [12] and [13] for adapting the mixture coefficients of SCHMM parameters and then extended to incremental adaptive learning of all of the CDHMM parameters in [14]. Based on the theory of recursive Bayesian inference, the QB algorithm is designed to incrementally update the hyperparameters on the approximate posterior distribution and the CDHMM parameters simultaneously [13], [14]. By further introducing some forgetting mechanisms, namely exponential forgetting and hyperparameter refreshing [14], to adjust the contribution of previously observed sample utterances, the algorithm is truly adaptive in nature and capable of performing an ideal on-line adaptive learning using only the current sample utterance. On the other hand, the QB framework is also flexible enough

Manuscript received October 22, 1996; revised June 11, 1997. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joseph Picone.

Q. Huo was with ATR Interpreting Telecommunications Research Laboratories, Kyoto 619-02, Japan. He is now with the Department of Computer Science and Information Systems, University of Hong Kong, Hong Kong (e-mail: qhuo@cs.hku.hk).

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

Publisher Item Identifier S 1063-6676(98)04215-1.

to include the batch and/or block mode MAP/ML learning as special cases [14].

In the above-mentioned HMM-based Bayesian adaptation framework, HMM parameters of different speech units are usually assumed independent. Therefore, each model can only be adapted if the corresponding speech unit has been observed in the current adaptation data. Consequently, only after all units have been observed enough times, all of the HMM parameters can thus be effectively adapted. To enhance the efficiency and the effectiveness of the Bayes adaptive learning, it is desirable to introduce some constraints on HMM parameters based on all possible sources of knowledge. Therefore, all the model parameters can be adjusted at the same time in a consistent and systematic way even though some units are not seen in adaptation data. A simple way to achieve the above objective is to introduce the parameter tying. Consequently, the formulation in [14] can be straightforwardly modified to accommodate the on-line adjustment of the tied parameters. Another way to achieve the above objective is to explicitly consider the correlation of HMM parameters corresponding to different speech units, and it is this kind of approach and strategy on which this work focuses. However, it is too difficult to define a joint prior distribution for all sets of HMM parameters, if not impossible. A tractable case could be to assume all mean vectors are correlated and have a joint prior distribution [18]. In this paper, we restrict ourselves to this special case and extend our QB learning framework to cope with the correlated CDHMM's with Gaussian mixture state observation densities in which all mean vectors are assumed to be correlated and have a joint Gaussian distribution. Considering the difficulties of parameter updating and initial hyperparameters' estimation arisen from the introduction of correlation between different models, we propose, in this paper, a successive approximation algorithm based on pairwise correlations to update the mean vectors of CDHMM's as well as the corresponding hyperparameters. As an example, the method is applied to on-line speaker adaptation and its viability is confirmed in a series of comparative experiments using a 26-letter English alphabet vocabulary.

The rest of the paper is organized as follows. After a brief introduction of the concept of the recursive Bayesian inference for CDHMM's, the QB formulation for incremental training of the correlated CDHMM's is presented in Section II. A successive approximation algorithm is proposed to implement the correlated mean vectors' updating and the resultant on-line adaptation algorithm is described in Section III. Some important implementation issues are discussed in Section IV. In Section V, a series of experimental results along with discussions and analyses for an incremental speaker adaptation application are reported. Finally, we summarize our findings in Section VI.

II. Quasi-Bayes Learning of Correlated CDHMM's

Consider a collection of M CDHMM's $\Lambda = \{\lambda_q\}_{q=1,\cdots,M}$, where $\lambda_q = (\pi^{(q)}, A^{(q)}, \theta^{(q)})$ denotes the set of parameters of the qth N-state CDHMM used to characterize the qth speech unit, of which $\pi^{(q)} = [\pi_1^{(q)}, \pi_2^{(q)}, \cdots, \pi_N^{(q)}]^t$ represents the initial state distribution, $A^{(q)} = [a_{ij}^{(q)}]$ is the transition probability matrix, and $\theta^{(q)}$ is the parameter vector composed of mixture parameters $\theta_i^{(q)} = \{\omega_{ik}^{(q)}, m_{ik}^{(q)}, \Sigma_{ik}^{(q)}\}$ for state i. The state observation probability density function (pdf) is assumed to be a mixture of multivariate Gaussian pdf's, as follows:

$$p(\mathbf{x} \mid \theta_i^{(q)}) = \sum_{k=1}^{K} \omega_{ik}^{(q)} \mathcal{N}(\mathbf{x} \mid m_{ik}^{(q)}, \Sigma_{ik}^{(q)}) \tag{1}$$

where the mixture coefficients $\omega_{ik}^{(q)}$'s satisfy the constraint $\sum_{k=1}^{K} \omega_{ik}^{(q)} = 1$, and $\mathcal{N}(\mathbf{x} \mid m_{ik}^{(q)}, \Sigma_{ik}^{(q)})$ is the kth normal mixture component with $m_{ik}^{(q)}$ being the D-dimensional mean vector and $\Sigma_{ik}^{(q)}$ being the $D \times D$ covariance matrix with its dth diagonal element being $\sigma_{ik}^{(q)2}(d)$. For notational convenience, it is assumed that all the state observation pdf's have the same number of mixture components.

Let $\mathcal{X}_1^n = \{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_n\}$ be n independent sets of observation samples that are used to estimate the CDHMM parameters Λ. Our initial knowledge about Λ is assumed to be contained in a known joint a priori density $p(\Lambda)$. Let us assume the samples $\mathcal{X}_i$'s are given successively one by one, we can obtain (see, e.g., [7]) a recursive expression for the a posteriori pdf of Λ, given $\mathcal{X}_1^n$, as

$$p(\Lambda \mid \mathcal{X}_1^n) = \frac{p(\mathcal{X}_n \mid \Lambda) \cdot p(\Lambda \mid \mathcal{X}_1^{n-1})}{\int_\Omega p(\mathcal{X}_n \mid \Lambda) \cdot p(\Lambda \mid \mathcal{X}_1^{n-1}) d\Lambda} \tag{2}$$

where Ω denotes an admissible region of the CDHMM parameter space. Starting the calculation of posterior pdf from $p(\Lambda)$, repeated use of the (2) produces the sequence of densities $p(\Lambda \mid \mathcal{X}_1^1)$, $p(\Lambda \mid \mathcal{X}_1^2)$, and so forth. This provides a basis of making formal recursive Bayesian inference of parameters Λ. However, there are some serious computational difficulties to directly implement this learning procedure [14]. Consequently, some approximations are needed in practice.

In this study, we only consider the case of CDHMM's in which the covariance matrices are specified. We define the parameter vector $\mathbf{m}$ to be the collection of the mean vectors of all the Gaussian mixture components of CDHMM's and denoted simply by an operator *vec* as $\mathbf{m} = vec\{m_{ik}^{(q)}\}$. We also define another operator *block-diag* to denote a block diagonal matrix, e.g., $\Xi = \textit{block-diag}\{\Sigma_{ik}^{(q)}\}$, with each diagonal block element to be also a matrix, e.g., $\Sigma_{ik}^{(q)}$. Further denote $\lambda_q' = (\pi_i^{(q)}, a_{ij}^{(q)}, \omega_{ik}^{(q)})$. Although in principle, any parametric form could be adopted for the initial prior pdf of Λ, a careful choice of it can lead to a more mathematically tractable solution of the problem. The key is the concept of conjugate prior distribution, and the reader is referred to [3], e.g., for more details on the idea. It is well known that there exist no natural conjugate densities for CDHMM 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 [10], [12], [14]. However, we still can benefit from assuming our initial prior pdf to have the same parametric form as the conjugate pdf of the complete-data density as shown in [10], [12], [14] and

[20]. Similarly here, the initial prior pdf of Λ is assumed to be

$$g(\Lambda) = g(\mathbf{m}) \prod_{q=1}^{M} g(\lambda_q') \tag{3}$$

where

$$g(\lambda_q') \propto \prod_{i=1}^{N} \left\{ \left[\pi_i^{(q)}\right]^{\eta_i^{(q)}-1} \cdot \left(\prod_{j=1}^{N} \left[a_{ij}^{(q)}\right]^{\eta_{ij}^{(q)}-1} \right) \cdot \left(\prod_{k=1}^{K} \left[\omega_{ik}^{(q)}\right]^{\nu_{ik}^{(q)}-1} \right) \right\} \tag{4}$$

takes the special form of a matrix beta pdf with sets of positive hyperparameters of $\{\eta_i^{(q)}\}, \{\eta_{ij}^{(q)}\}, \{\nu_{ik}^{(q)}\}$ [10], [12], [14], and

$$g(\mathbf{m}) = \mathcal{N}(\mathbf{m} \mid \boldsymbol{\mu}, \mathbf{U}) \tag{5}$$

has a joint normal pdf with mean vector $\boldsymbol{\mu} = vec\{\mu_{ik}^{(q)}\}$ and covariance matrix $\mathbf{U}$ [18]. This class of prior distributions actually constitutes a conjugate family of the complete-data density and is denoted as $\mathcal{P}$. In (4), "$\propto$" denotes proportionality.

The quasi-Bayes procedure is, at each step of the recursive Bayes learning, to approximate the true posterior distribution $p(\Lambda \mid \mathcal{X}_1^n)$, by the "closest" tractable distribution $g(\Lambda \mid \varphi^{(n)})$ within the given class $\mathcal{P}$, under the criterion of both distributions having the same (local) mode [14]. Here $\varphi^{(n)}$ denotes the updated hyperparameters after observing the samples $\mathcal{X}_n$. More specifically, consider at time instant n, we have a training set $\mathcal{X}_n = \{\mathbf{x}_n^{(q,r)}\}$ and our prior knowledge about Λ is approximated by $g(\Lambda \mid \varphi^{(n-1)})$. Here $\mathbf{x}_n^{(q,r)}$ denotes the rth training observation sequence of length $T_n^{(q,r)}$ associated with the qth speech unit, and each unit has $W_q^{(n)}$ such observation sequences. Let $\mathcal{Y}_n = (\mathcal{X}_n, \mathcal{Z}_n)$ denote the associated complete-data and $\mathcal{Z}_n = \{\mathbf{s}_n^{(q,r)}, \mathbf{l}_n^{(q,r)}\}$ be corresponding missing-data, where $\mathbf{s}_n^{(q,r)}$ denotes the unobserved state sequence and $\mathbf{l}_n^{(q,r)}$ is the sequence of the unobserved mixture component labels corresponding to the observation sequence $\mathbf{x}_n^{(q,r)}$. Given the set of observation sequences $\{\mathbf{x}_n^{(q,r)}\}$ and the above prior pdf $g(\Lambda \mid \varphi^{(n-1)})$, we can get the approximate MAP estimate $\Lambda^{(n)}$ of Λ by repeating the following EM steps.

E-step: Compute

$$R(\Lambda \mid \Lambda^{(n-1,l-1)}) = \kappa \cdot \log g(\Lambda \mid \varphi^{(n-1)}) + E[\log p(\mathcal{Y}_n \mid \Lambda) \mid \mathcal{X}_n, \Lambda^{(n-1,l-1)}] \tag{6}$$

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

M-step: Choose

$$\Lambda^{(n-1,l)} = \underset{\Lambda}{\operatorname{argmax}}\, R(\Lambda \mid \Lambda^{(n-1,l-1)}) \tag{7}$$

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

By choosing the initial prior pdf to be the conjugate family of the complete-data density, it can be verified that with an appropriate normalization factor C, $C \cdot \exp\{R(\Lambda \mid \Lambda^{(n-1,l-1)})\}$ 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:

$$\hat{\eta}_i^{(q)} = \kappa \cdot \left(\eta_i^{(q)} - 1\right) + 1 + \sum\nolimits_{r=1}^{W_q} \gamma_1^{(q,r)}(i) \tag{8}$$

$$\hat{\eta}_{ij}^{(q)} = \kappa \cdot \left(\eta_{ij}^{(q)} - 1\right) + 1 + \sum\nolimits_{r=1}^{W_q} \sum\nolimits_{t=1}^{T^{(q,r)}} \gamma_t^{(q,r)}(i,j) \tag{9}$$

$$\hat{\nu}_{ik}^{(q)} = \kappa \cdot \left(\nu_{ik}^{(q)} - 1\right) + 1 + c_{ik}^{(q)} \tag{10}$$

$$\hat{\boldsymbol{\mu}} = \kappa\Xi(\kappa\Xi + \mathbf{UC})^{-1}\boldsymbol{\mu} + \mathbf{U}(\kappa\Xi + \mathbf{CU})^{-1}\mathbf{C}\bar{\mathbf{X}} \tag{11}$$

$$\hat{\mathbf{U}} = \mathbf{U}(\kappa\Xi + \mathbf{CU})^{-1}\Xi \tag{12}$$

where

$$\gamma_t^{(q,r)}(i,j) = \Pr\left(s_t^{(q,r)} = i, s_{t+1}^{(q,r)} = j \mid \mathbf{x}^{(q,r)}, \Lambda\right) \tag{13}$$

$$\gamma_t^{(q,r)}(i) = \Pr\left(s_t^{(q,r)} = i \mid \mathbf{x}^{(q,r)}, \Lambda\right) \tag{14}$$

$$\zeta_t^{(q,r)}(i,k) = \Pr\left(s_t^{(q,r)} = i, l_t^{(q,r)} = k \mid \mathbf{x}^{(q,r)}, \Lambda\right) \tag{15}$$

and these terms can be computed efficiently by using the forward-backward algorithm (e.g., [16], [25]). Further

$$\mathbf{C} = \textit{block-diag}\left\{c_{ik}^{(q)} \cdot I_{D\times D}\right\} \tag{16}$$

$$\bar{\mathbf{X}} = vec\left\{\bar{\mathbf{x}}_{ik}^{(q)}\right\} \tag{17}$$

with

$$c_{ik}^{(q)} = \sum\nolimits_{r=1}^{W_q} \sum\nolimits_{t=1}^{T^{(q,r)}} \zeta_t^{(q,r)}(i,k) \tag{18}$$

$$\bar{\mathbf{x}}_{ik}^{(q)} = \sum\nolimits_{r=1}^{W_q} \sum\nolimits_{t=1}^{T^{(q,r)}} \zeta_t^{(q,r)}(i,k) \cdot \mathbf{x}_t^{(q,r)} / c_{ik}^{(q)} \tag{19}$$

and $I_{D\times D}$ is an identity matrix. Note that for notational simplicity, we've 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 $g(\Lambda \mid \hat{\varphi})$ and are shown as follows:

$$\hat{\pi}_i^{(q)} = \frac{\hat{\eta}_i^{(q)} - 1}{\sum_{j=1}^{N}\left(\hat{\eta}_j^{(q)} - 1\right)} = \frac{\kappa \cdot \left(\eta_i^{(q)} - 1\right) + \sum_{r=1}^{W_q} \gamma_1^{(q,r)}(i)}{\sum_{j=1}^{N}\left[\kappa \cdot \left(\eta_j^{(q)} - 1\right) + \sum_{r=1}^{W_q} \gamma_1^{(q,r)}(j)\right]} \tag{20}$$

$$\hat{a}_{ij}^{(q)} = \frac{\hat{\eta}_{ij}^{(q)} - 1}{\sum_{k=1}^{N}\left(\hat{\eta}_{ik}^{(q)} - 1\right)} = \frac{\kappa \cdot \left(\eta_{ij}^{(q)} - 1\right) + \sum_{r=1}^{W_q} \sum_{t=1}^{T^{(q,r)}} \gamma_t^{(q,r)}(i,j)}{\sum_{k=1}^{N}\left[\kappa \cdot \left(\eta_{ik}^{(q)} - 1\right) + \sum_{r=1}^{W_q} \sum_{t=1}^{T^{(q,r)}} \gamma_t^{(q,r)}(i,k)\right]} \tag{21}$$

$$\hat{\omega}_{ik}^{(q)} = \frac{\hat{\nu}_{ik}^{(q)} - 1}{\sum_{j=1}^{K}\left(\hat{\nu}_{ij}^{(q)} - 1\right)} = \frac{\kappa \cdot \left(\nu_{ik}^{(q)} - 1\right) + \sum_{r=1}^{W_q} \sum_{t=1}^{T^{(q,r)}} \zeta_t^{(q,r)}(i,k)}{\sum_{j=1}^{K}\left[\kappa \cdot \left(\nu_{ij}^{(q)} - 1\right) + \sum_{r=1}^{W_q} \sum_{t=1}^{T^{(q,r)}} \zeta_t^{(q,r)}(i,j)\right]} \tag{22}$$

$$\hat{\mathbf{m}} = \hat{\boldsymbol{\mu}}. \tag{23}$$

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

$$p(\Lambda \mid \mathcal{X}_n) = \frac{p(\mathcal{X}_n \mid \Lambda) \cdot g(\Lambda \mid \varphi^{(n-1)})}{\int_\Omega p(\mathcal{X}_n \mid \Lambda) \cdot g(\Lambda \mid \varphi^{(n-1)}) d\Lambda}. \tag{24}$$

Thus, the hyperparameters $\varphi^{(n)}$ are obtained at the last (actually Lth) EM iteration by using the equations in (8)–(12) to satisfy

$$g(\Lambda \mid \varphi^{(n)}) \propto \exp\{R(\Lambda \mid \Lambda^{(n-1,L-1)})\} \tag{25}$$

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

The above forward-backward type procedure can be easily extended to a segmental (or Viterbi) one by replacing (13)–(15) with

$$\gamma_t^{(q,r)}(i,j) = \delta(s_t^{(q,r)} - i)\delta(s_{t+1}^{(q,r)} - j) \tag{26}$$

$$\gamma_t^{(q,r)}(i) = \delta(s_t^{(q,r)} - i) \tag{27}$$

$$\zeta_t^{(q,r)}(i,k) = \gamma_t^{(q,r)}(i) \cdot \frac{\omega_{ik}^{(q)} \mathcal{N}(\mathbf{x}_t^{(q,r)} \mid m_{ik}^{(q)}, \Sigma_{ik}^{(q)})}{\sum_{j=1}^K \omega_{ij}^{(q)} \mathcal{N}(\mathbf{x}_t^{(q,r)} \mid m_{ij}^{(q)}, \Sigma_{ij}^{(q)})} \tag{28}$$

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

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

III. Successive Approximation Based on Pairwise Correlation

A. Algorithm

Suppose that we only have the knowledge of pairwise correlations between different mean vectors instead of trying to exploit the joint correlation structure of all the mean vectors. Further suppose that we do not consider the correlation between different dimensional elements of the same mean vector or two different mean vectors. Actually, the second assumption is not necessary and the succeeding formulation can be easily extended to cope with the more general cases. We use this assumption in the following discussion i) for the simplicity of the description of the algorithm and ii) because we use this simple formulation in our experiments. With the above two assumptions, we can simplify the following discussion to a one-dimensional (1-D) case. So, every time, we only need to consider a pair of random variables $m_{ikd}^{(q)}$ and $m_{i'k'd}^{(q')}$. For notational simplicity, they are denoted, respectively as $m_I(d)$ and $m_{I'}(d)$. We assume $m_I(d)$ and $m_{I'}(d)$ have a joint a priori normal pdf with means $\mu_I(d)$ and $\mu_{I'}(d)$, variances $u_I^2(d)$ and $u_{I'}^2(d)$, and covariance $\rho_{II'}(d) \cdot u_I(d) \cdot u_{I'}(d)$, where $\rho_{II'}(d)$ is the correlation coefficient. We pretend only $c_I = c_{ik}^{(q)}$ observations belonging to m_I are obtained and no observations for $m_{I'}$ are available. Given these observations, it can be shown by using the (11) and (12) that the joint posterior pdf of $m_I(d)$ and $m_{I'}(d)$ is still a normal one with the following hyperparameters:

$$\tilde{\mu}_I(d) = \frac{\kappa\sigma_I^2(d)}{\kappa\sigma_I^2(d) + c_I u_I^2(d)} \mu_I(d) + \frac{c_I u_I^2(d)}{\kappa\sigma_I^2(d) + c_I u_I^2(d)} \bar{x}_I(d) \tag{29}$$

$$= \mu_I(d) + \frac{c_I u_I^2(d)}{\kappa\sigma_I^2(d) + c_I u_I^2(d)} (\bar{x}_I(d) - \mu_I(d)) \tag{30}$$

$$\tilde{\mu}_{I'}(d) = \mu_{I'}(d) + \frac{c_I \rho_{II'}(d) u_I(d) u_{I'}(d)}{\kappa\sigma_I^2(d) + c_I u_I^2(d)} (\bar{x}_I(d) - \mu_I(d)) \tag{31}$$

$$= \mu_{I'}(d) + \frac{\rho_{II'}(d) u_{I'}(d)}{u_I(d)} (\tilde{\mu}_I(d) - \mu_I(d)) \tag{32}$$

$$\tilde{u}_I^2(d) = \frac{\sigma_I^2(d)}{\kappa\sigma_I^2(d) + c_I u_I^2(d)} u_I^2(d) \tag{33}$$

$$\tilde{u}_{I'}^2(d) = \frac{\kappa\sigma_I^2(d) + c_I u_I^2(d)\big(1 - \rho_{II'}^2(d)\big)}{\kappa\big(\kappa\sigma_I^2(d) + c_I u_I^2(d)\big)} u_{I'}^2(d) \tag{34}$$

$$\tilde{\rho}_{II'}(d) = \frac{\rho_{II'}(d)}{\sqrt{1 + \frac{c_I u_I^2(d)}{\kappa\sigma_I^2(d)}\big(1 - \rho_{II'}^2(d)\big)}}. \tag{35}$$

If we define $\tau_I(d) = \sigma_I^2(d)/u_I^2(d)$, the above equations become

$$\tilde{\mu}_I(d) = \frac{\kappa\tau_I(d)}{\kappa\tau_I(d) + c_I} \mu_I(d) + \frac{c_I}{\kappa\tau_I(d) + c_I} \bar{x}_I(d) \tag{36}$$

$$= \mu_I(d) + \frac{c_I}{\kappa\tau_I(d) + c_I} (\bar{x}_I(d) - \mu_I(d)) \tag{37}$$

$$\tilde{\mu}_{I'}(d) = \mu_{I'}(d) + \rho_{II'}(d) \frac{\sigma_{I'}(d)}{\sigma_I(d)} \times \sqrt{\frac{\tau_I(d)}{\tau_{I'}(d)}} \frac{c_I}{\kappa\tau_I(d) + c_I} (\bar{x}_I(d) - \mu_I(d)) \tag{38}$$

$$= \mu_{I'}(d) + \rho_{II'}(d) \frac{\sigma_{I'}(d)}{\sigma_I(d)} \sqrt{\frac{\tau_I(d)}{\tau_{I'}(d)}} (\tilde{\mu}_I(d) - \mu_I(d)) \tag{39}$$

[1] Strictly speaking, EM algorithm [4] can only guarantee the mode of the approximate pdf to approach a local maximum of the above true posterior pdf.

1. Estimate initial hyperparameters (details given in the next section). Set up (initial) top $\mathcal{K}$ prediction tables (explained in the following section).
2. Receive (an) utterance(s) to be recognized.
3. Do acoustic normalization/equalization as required.
4. Do recognition and record results.
5. Do supervised (if permitted) or unsupervised incremental adaptation as follows:
   - in case of changeable top $\mathcal{K}$ tables, update them based on current correlation coefficients; otherwise, skip this step.
   - do EM-iterations as follows:
     - initialize hyperparameters to be the latest history ones.
     - for those speech unit having observation data
       - update state transition matrices.
       - update mixture coefficients.
     - update mean vectors with successive approximation algorithm as follows:
       - reset temporary hyperparameters.
       - choose a mixture component "I" having observation data but not processed
         - identify top $\mathcal{K}$ mixture components "I'"'s most correlated to mixture component I
         - for each mixture component I', update its temporary hyperparameters as in equations (31), (34) and (35).
         - update temporary hyperparameters for mixture component I as in equations (30) and (33)
       - if all the mixture components having observation data have been processed, go to next substep; otherwise, go back to previous substep.
       - update all mean vectors and exit the successive approximation algorithm.
   - update all hyperparameters.
6. Go to Step 2.

Fig. 1. On-line adaptation algorithm for correlated CDHMM's.

$$\tilde{\tau}_I(d) = \kappa\tau_I(d) + c_I \tag{40}$$

$$\tilde{\tau}_{I'}(d) = \frac{\kappa(\kappa\tau_I(d) + c_I)}{\kappa\tau_I(d) + c_I\big(1 - \rho^2_{II'}(d)\big)}\tau_{I'}(d) \tag{41}$$

$$\tilde{\rho}_{II'}(d) = \frac{\rho_{II'}(d)}{\sqrt{1 + \frac{c_I}{\kappa\tau_I(d)}\big(1 - \rho^2_{II'}(d)\big)}}. \tag{42}$$

By successively changing the role of the mixture component and repeating the above steps, we can approximately approach the updating of the hyperparameters in the (11) and (12). We then naturally come up with the on-line adaptation algorithm for correlated CDHMM's as shown in Fig. 1.

B. Discussion

Now, we are ready to compare our approach to other related methods in the literature. In the speech and pattern recognition area, to our knowledge, it was Lasry and Stern that first proposed a formulation of the MAP estimate (called extended MAP, or EMAP) in [18] for the mean vectors of a set of Gaussian pdf's in which those mean vectors are assumed to have a joint Gaussian prior distribution. They applied the EMAP method to the dynamic speaker adaptation in a feature-based isolated word recognition application [32]. To avoid the difficulty of the initial hyperparameters estimation, a classifier with a decision-tree structure is adopted. At each node of the decision tree, the utterance is classified into a small number of decision categories, based on a relatively small number of features that are relevant to the classification in question. Consequently, every time, they only make use of the correlation information among a small number of classes for adaptation and thus can afford the memory requirement and the computational complexity of the related algebraic operations. To avoid the repeated inversion of a big matrix in the standard EMAP implementation for dynamic speaker adaptation, later, in the context of SCHMM, Rozzi and Stern developed a least mean square (LMS) algorithm to implement the correlated means adaptation, which is supposed to be more computationally efficient, but at the expense of a finite misadjustment

[28]. On the other hand, the initial hyperparameters estimation problem still exists. More recently, Zavaliagkos *et al.* applied EMAP into a large scale CDHMM-based speech recognition systems [36], [37]. With a similar motivation as in [18] and [32], they adopted a hierarchical class tying technique to ease the above-mentioned difficulties of the EMAP implementation. In this study, we integrate EMAP into our quasi-Bayes learning framework and propose the above successive approximation algorithm to ease the implementation. The algorithm does not involve any big matrix operation, thus becomes very computationally efficient. On the other hand, even if we can have an initial estimate of a nonsingular matrix $\mathbf{U}$, the successive approximation algorithm cannot guarantee its nonsingularity after each iteration. However, because the implementation of the algorithm does not rely on the assumption of the nonsingularity of the matrix $\mathbf{U}$, this in turn eases the problem of the initial hyperparameters' estimation, as discussed in the next section.

We can also set up the links between our approach and two other techniques, namely MAP/VFS (e.g., [11], [24], [33]–[35]) and regression-based model prediction (RMP) methods (e.g., [1], [2], [8]). It is easy to verify that the (32) is very similar to the so-called interpolation step in MAP/VFS method [33], [34] except that i) we use a different weighting coefficient, and ii) every time, we only use the information from one mixture component to predict the mean vector of the mixture component without observations. But in our approach, by successively changing the role of the mixture components, we can achieve the similar effects as those of both interpolation and smoothing steps in MAP/VFS formulation. Furthermore, by updating the correlation coefficient as in (35), the algorithm can autonomously control the importance of the correlation information and thus make the estimations of the mean vectors of CDHMM asymptotically converge to their MAP or ML estimates without considering correlation. On the other hand, in MAP/VFS case, to avoid early saturation of the adaptation, some heuristic methods have to be employed [35]. We can also view (32) as a simple linear regression function with one explanatory variable $\tilde{\mu}_I(d)$ and the adaptive regression coefficients. Once again, by successive approximation, we can achieve the similar effect as that of RMP in [1] and [2].

We also wish to draw the reader's attention to the work of Shahshahani [30], who has a very similar motivation to our work in the sense of exploiting model correlations for efficient Bayesian adaptation where a Gibbs distribution is adopted to serve as the joint prior pdf of the mean vectors of the all CDHMM's. However, in that work, only conventional batch mode adaptation is formulated and it is very difficult to extend this method for a true on-line adaptive learning.

In the context of efficient adaptation, our method also shares the similarity with another type of transformation-based adaptation methods (e.g., [5], [23]) in a more general sense of global mapping. The basic idea of both types of methods is to bind HMM parameters together (via correlation structure in our case and some shared transformations among different model parameters in the latter case), and then to adjust them globally in a consistent and systematic way. For the transformation-based approaches, in order to achieve a better asymptotic convergence, one has to either dynamically increase the number of shared transformations according to the amount of available adaptation data (e.g., [23]) or just combined with the Bayesian approach (e.g., [6]), both in a heuristic way.

From above discussions, we can see that the Bayesian learning procedure suggested in this study has a more consistent formulation as well as an intuitively pleasing behavior (an improved adaptation efficiency for short adaptation data and a good asymptotic property for increasing number of adaptation data). By activating the forgetting mechanism, the algorithm can also be used to cope with the continuously changing conditions [14]. In the following sections, we will show by a series of experiments that the proposed algorithm does work and 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

We use a modified method of moment to estimate the initial correlation coefficients $\rho_{II'}^{(0)}(d)$ as in (43), shown at the bottom of the page, where $m_I^{(i)}$ is the ith set of mean vectors, $c_I^{(i)}$ is the corresponding "EM count," and $\bar{m}_I$ is the average of $m_I^{(i)}$'s. In the following speaker adaptation (SA) experiments, we use speaker-independent (SI) trained parameters to replace $\bar{m}_I$, and $m_I^{(i)}$ correspond to the parameters estimated from ith speaker or speaker group [12]. Apart from correlation coefficients, other initial hyperparameters are estimated as follows (for a detailed discussion, see [14]):

$$\eta_i^{(0)} = 1 + \epsilon_1 \cdot \gamma_1^{(\mathrm{SI})}(i) \tag{44}$$

$$\eta_{ij}^{(0)} = 1 + \epsilon_1 \cdot \sum\nolimits_t \gamma_t^{(\mathrm{SI})}(i,j) \tag{45}$$

$$\nu_{ik}^{(0)} = 1 + \epsilon_1 \cdot \sum\nolimits_t \zeta_t^{(\mathrm{SI})}(i,k) \tag{46}$$

$$\mu_{ik}^{(0)} = m_{ik}^{(\mathrm{SI})} \tag{47}$$

$$u_{ikd}^{2(0)} = \sigma_{ikd}^{2(\mathrm{SI})} \cdot \left[\epsilon_1 \cdot \sum\nolimits_t \zeta_t^{(\mathrm{SI})}(i,k)\right]^{-1} \tag{48}$$

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.

$$\rho_{II'}^{(0)}(d) = \frac{\sum_i c_I^{(i)}\left(m_I^{(i)}(d) - \bar{m}_I(d)\right) c_{I'}^{(i)}\left(m_{I'}^{(i)}(d) - \bar{m}_{I'}(d)\right)}{\sqrt{\sum_i c_I^{(i)2}\left(m_I^{(i)}(d) - \bar{m}_I(d)\right)^2} \cdot \sqrt{\sum_i c_{I'}^{(i)2}\left(m_{I'}^{(i)}(d) - \bar{m}_{I'}(d)\right)^2}} \tag{43}$$

B. Top $\mathcal{K}$ Prediction and Possible Constraints

For each mixture component having observation data, we only use its observed information to predict other $\mathcal{K}$ mixture components which have the highest top $\mathcal{K}$ values based on, among many possibilities, the following two between-component correlation measures [1], [2]:

$$\bar{\rho}_{II'} = \frac{1}{D}\sum_{d=1}^{D}|\rho_{II'}(d)| \tag{49}$$

or

$$\bar{\rho}_{II'} = \sqrt{\frac{1}{D}\sum_{d=1}^{D}\rho_{II'}^2(d)}. \tag{50}$$

In the following experiments, we only consider the correlation of mixture components between different speech units. If we allow the correlations between the mixture components of the same state, or neighboring states, this will have the similar effects of Zhao's so-called context-modulation [38]. We do not use the correlation between the mixture components belonging to the same basic unit to avoid oversmoothing. Based on these constraints, we can set up a top $\mathcal{K}$ prediction table for each mixture component. These tables can be either fixed during on-line adaptation or dynamically changed based on the updated correlation coefficients. Further constraints can also be applied to limit the correlated mixture components' domain based on some acoustic-phonetic knowledge (e.g., only consider the correlation between different speech units with a similar acoustic nature) and/or some data-driven clustering results. We will not further investigate these engineering issues here and leave them for future study.

V. Speaker Adaptation Experiments

A. Experimental Setup

To examine its viability, the proposed algorithm 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 data bases 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 SI training and initial prior density estimation, the OGI ISOLET data base 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. The remaining 16 tokens were collected in eight different sessions in which two tokens of each letter were collected in each session. For each person and each letter, we divide equally those 16 tokens collected in eight different sessions into two parts, one for adaptive training, another for testing.

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 [19]. Although there are other alternatives (e.g., [29]), only utterance-based cepstral mean subtraction (CMS) was applied for acoustic normalization. In recognition, the decision rule determined the recognized letter as the one which attained the highest forward-backward probability.

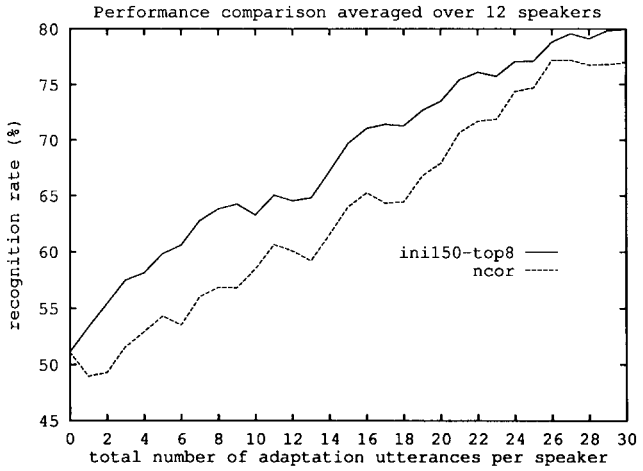
In the following sections, we study the convergence property of the algorithm, the effects of top $\mathcal{K}$ prediction, the effects of different initial hyperparameters estimations, the effects of different number of EM iterations, and finally the comparison between forward-backward and segmental QB learning. Except explicitly stated, in most of the experiments, we use the following default setup:

1) initial hyperparameters $\rho_{II'}^{(0)}(d)$'s estimated from 150 sets of speaker-dependent (SD) models;
2) between-component correlation measure in (49);
3) forward-backward type QB procedure;
4) three EM iterations for on-line adaptation;
5) fixed top $\mathcal{K}$ prediction tables.

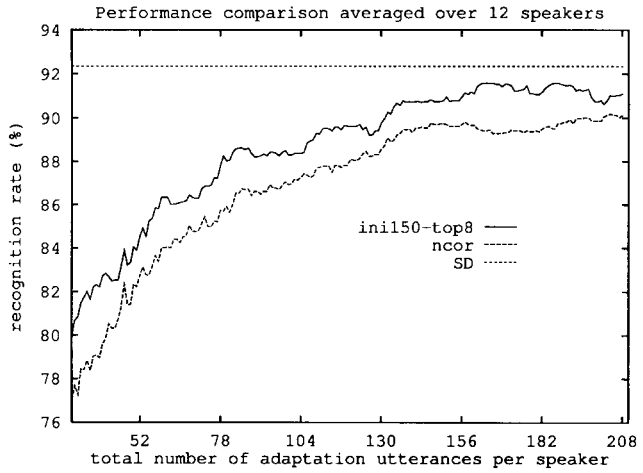
All of the experiments were performed in a supervised mode and no forgetting mechanism is activated.

B. Convergence Property

Starting with a set of SI initial models, we present training tokens for each letter cyclically and perform utterance-based supervised on-line adaptation (OLA). After each OLA step, we test the recognizer on a separate testing set to measure the performance changes. We plot in Fig. 2(a) and (b) the performance (word accuracy in %) comparison of two OLA setups, averaged over 12 speakers, as a function of total number of adaptation tokens per speaker. In these figures, "ncor" stands for the OLA experiment without considering correlation between mixture components. "ini150-top8" refers to the case of considering top eight mixture components prediction. "SD" refers to the recognition performance averaged over 12 speakers by using SD models trained from eight adaptation tokens per letter for each speaker. Fig. 2(a) shows the fast adaptation effects while Fig. 2(b) checks the asymptotic property of the algorithm. The experimental results show that the proposed algorithm improves the OLA performance further by considering the correlation information and also has a good asymptotic convergence behavior. In the particular experiments here, the results show that the SA performance



(a)



(b)

Fig. 2. Performance (word accuracy in %) comparison as a function of total number of adaptation tokens per speaker. (a) Fast adaptation effect. (b) Asymptotic convergence property (three EM iterations, SD recognition rate is 92.3%).

is inferior to the SD performance when enough SD training tokens are available (here eight tokens per letter).

C. Effects of Top $\mathcal{K}$ Prediction

We will examine experimentally three issues related to top $\mathcal{K}$ prediction, namely how to define the neighborhood, how to choose the size (value of $\mathcal{K}$) of the neighborhood, and finally fixed versus dynamically defined neighborhood. In Section IV, we discussed two methods for deriving top $\mathcal{K}$ prediction table. In Table I, we compare the OLA performance by using those two between-component correlation measures. The experimental results show that the measure in (49) [2] (denoted as "abs") performs slightly better than the one in (50) [1] (denoted as "sqr") in our specific experimental conditions here. So, in the remaining experiments, we use the first measure.

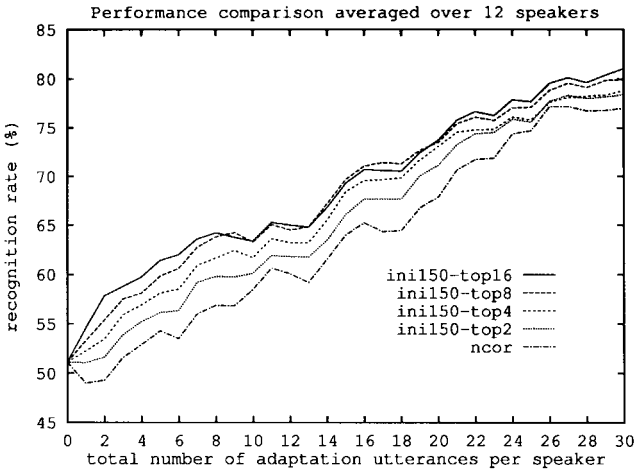
In Figs. 3(a) and (b), we compare the effects of different $\mathcal{K}$ on the OLA performance improvement. Intuitively, with less adaptation data, it will be helpful to rely more on correlation information (a bigger $\mathcal{K}$) to enhance the OLA effect. On the other hand, when more adaptation data become available, we should rely less on the correlation information (a smaller $\mathcal{K}$), and thus a better asymptotic convergence is expected. The results shown in Fig. 3 confirmed this expectation. So, in practice, to achieve a better performance, a possible strategy could be to dynamically redefine and/or to shrink (e.g., via decreasing the $\mathcal{K}$) the neighborhood while the number of total adaptation tokens increases. We will not go further here about the latter and leave this for future research.

TABLE I
PERFORMANCE (WORD ACCURACY IN %) COMPARISON AVERAGED OVER 12 SPEAKERS AS A FUNCTION OF TOTAL NUMBER OF ADAPTATION TOKENS PER SPEAKER BY USING DIFFERENT BETWEEN-COMPONENT CORRELATION MEASURES FOR TOP $\mathcal{K}$ PREDICTION TABLE SETUP (THREE EM ITERATIONS, $\mathcal{K} = 8$)

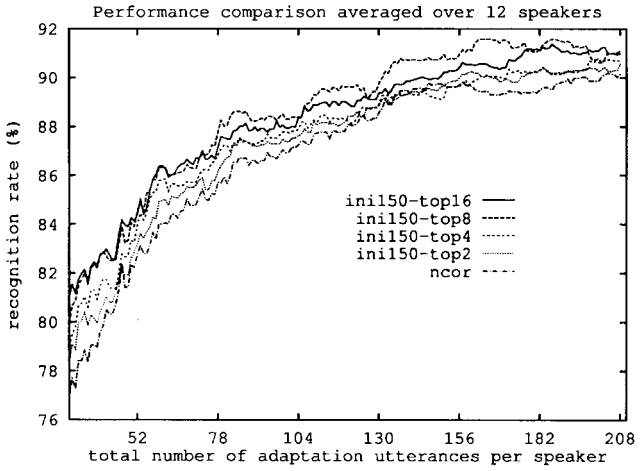
total number of adaptation tokens	on-line adaptation methods		
	abs	sqr	ncor
0	51.10	51.10	51.10
5	59.84	59.80	54.31
10	63.28	63.89	58.48
15	69.70	69.50	64.01
20	73.51	73.59	67.90
26	78.84	78.48	77.19
52	84.53	84.65	82.76
78	87.78	86.86	85.73
104	88.38	88.30	87.25
130	89.78	89.34	88.58
156	90.94	90.54	89.70
182	91.18	90.86	89.50
208	91.10	90.66	90.14

To examine the effect of the dynamically defined neighborhood, we report in Table II the performance comparison between fixed (denoted as "fix") and dynamically defined (denoted as "dynamic") neighborhood methods. The fixed-neighborhood method means that the top $\mathcal{K}$ prediction table is derived from the initial correlation coefficients and will be fixed during the adaptation process, albeit the correlation coefficients are updated. On the other hand, in the dynamically defined neighborhood method, we will dynamically update the top $\mathcal{K}$ prediction table based on the updated correlation coefficients. Our experimental results show that the former achieves a better asymptotic performance. This is not surprising when the following fact is considered. In our simple method to update the neighborhood, we rely on the updated correlation coefficients whose values will decrease while more adaptation data become available and this might lead to a less meaningful and unreliable neighborhood definition. On the other hand, we can get a more stable or robust result to use the initial correlation coefficients to define the neighborhood. However, we expect that a better strategy to dynamically define the neighborhood will eventually be helpful to enhance the OLA performance.

The above top $\mathcal{K}$ prediction related issues are important for efficient on-line adaptation. As a first step, we only use



(a)



(b)

Fig. 3. Performance (word accuracy in %) comparison as a function of total number of adaptation tokens per speaker by using different values of $\mathcal{K}$ for top $\mathcal{K}$ prediction. (a) Fast adaptation effect. (b) Asymptotic convergence property (three EM iterations).

the pure data-driven method and very simple constraints. We expect that discovering an appropriate acoustic space configuration and a good definition of an appropriate correlation structure among states and/or phones could be helpful for enhancing the efficiency of the OLA of the correlated CDHMM's. It will be interesting to see how it works by combining our approach with other techniques such as tree-structured Gaussians to explore acoustic space structure (e.g., [31]), phone-dependence tree to explore phonetic dependency structure (e.g., [27]), and their combination (e.g., [15]). We believe this is an area that deserves a further research from both a theoretical and a practical point of view.

D. Effects of Different Initial Hyperparameters Estimations

In Table III, we compare the effects of different estimations of hyperparameters $\rho_{II'}^{(0)}(d)$'s on the OLA performance improvement. We report the results on two cases of hyperparameters estimation where fixed top $\mathcal{K}$ prediction tables are used. "ini150" stands for the case in which a set of CDHMM's are trained for each speaker, and these 150 sets of SD models

TABLE II
PERFORMANCE (WORD ACCURACY IN %) COMPARISON AVERAGED OVER 12 SPEAKERS AS A FUNCTION OF TOTAL NUMBER OF ADAPTATION TOKENS PER SPEAKER BETWEEN FIXED AND DYNAMICALLY DEFINED TOP $\mathcal{K}$ PREDICTION TABLES (THREE EM ITERATIONS, $\mathcal{K} = 8$)

total number of adaptation tokens	on-line adaptation methods		
	fix	dynamic	ncor
0	51.10	51.10	51.10
5	59.84	59.92	54.31
10	63.28	62.92	58.48
15	69.70	69.70	64.01
20	73.51	73.75	67.90
26	78.84	78.12	77.19
52	84.53	83.49	82.76
78	87.78	86.85	85.73
104	88.38	87.81	87.25
130	89.78	89.10	88.58
156	90.94	89.98	89.70
182	91.18	90.66	89.50
208	91.10	91.02	90.14

(possibly poor estimations due to the insufficient training data) are used for hyperparameters estimation in (43). "ini16" refers to the case in which we first cluster 150 speakers into 16 groups [12], then train a set of CDHMM's for each speaker group, and finally use these 16 sets of models (supposedly better estimations due to more training data for each model) for hyperparameters estimation in (43). Although only two tokens for each letter are available in the former case, because we are using "EM count" as a weighting coefficient to automatically take into account the reliability of the HMM parameters estimation, we are getting better results by using more samples (albeit possibly a poor estimation for each sample) in the moment estimate of the correlation coefficients.

E. Effects of Different Number of EM Iterations

In Table IV, we compare the effects of different number of EM iterations on the OLA performance improvement. "3-EM" stands for the case where three EM iterations are performed for each OLA step, and similarly, "2-EM" and "1-EM" correspond to the cases of two and one EM iterations respectively. We found that there is no big difference of the performance by performing different number of EM iterations.

F. Forward-Backward versus Segmental QB Learning

In Table V, we compare the effects of forward-backward type and segmental type QB learning on the OLA performance improvement. "fb" stands for the case where forward-backward QB algorithm is used and three EM iterations are performed for OLA, and "seg" refers to its segmental counterpart. We found that there is no big difference of the performance between these two procedures. The similar fact is also observed in cases with fewer EM iterations (one

TABLE III
PERFORMANCE (WORD ACCURACY IN %) COMPARISON AVERAGED OVER 12 SPEAKERS AS A FUNCTION OF TOTAL NUMBER OF ADAPTATION TOKENS PER SPEAKER BY USING DIFFERENT ESTIMATIONS OF HYPERPARAMETERS $\rho_{ll'}^{(0)}(d)$'s (3 EM ITERATIONS, $\mathcal{K} = 8$)

total number of adaptation tokens	on-line adaptation methods		
	ini150	ini16	ncor
0	51.10	51.10	51.10
5	59.84	56.71	54.31
10	63.28	59.76	58.48
15	69.70	65.49	64.01
20	73.51	70.30	67.90
26	78.84	78.64	77.19
52	84.53	84.33	82.76
78	87.78	86.85	85.73
104	88.38	88.38	87.25
130	89.78	89.30	88.58
156	90.94	90.70	89.70
182	91.18	90.98	89.50
208	91.10	91.34	90.14

TABLE IV
PERFORMANCE (WORD ACCURACY IN %) COMPARISON AVERAGED OVER 12 SPEAKERS AS A FUNCTION OF TOTAL NUMBER OF ADAPTATION TOKENS PER SPEAKER BY PERFORMING DIFFERENT NUMBER OF EM ITERATIONS DURING ON-LINE ADAPTATION ($\mathcal{K} = 8$)

total number of adaptation tokens	on-line adaptation methods			
	3-EM	2-EM	1-EM	ncor
0	51.10	51.10	51.10	51.10
5	59.84	60.04	60.80	54.31
10	63.28	63.60	64.09	58.48
15	69.70	69.46	70.06	64.01
20	73.51	73.39	73.35	67.90
26	78.84	78.96	78.32	77.19
52	84.53	84.77	84.81	82.76
78	87.78	87.97	87.22	85.73
104	88.38	89.06	88.94	87.25
130	89.78	90.22	90.26	88.58
156	90.94	91.06	90.82	89.70
182	91.18	91.26	91.02	89.50
208	91.10	91.34	91.38	90.14

and two). The observations in this and previous subsections are especially meaningful for real-time implementation of the proposed algorithm in real applications. We can thus suggest with certain confidence that in applications where computational complexity is one of the main concerns, the segmental QB learning procedure with one EM iteration for each OLA step will provide a satisfactory solution in most of the cases.

TABLE V
PERFORMANCE (WORD ACCURACY IN %) COMPARISON AVERAGED OVER 12 SPEAKERS AS A FUNCTION OF TOTAL NUMBER OF ADAPTATION TOKENS PER SPEAKER BETWEEN FORWARD-BACKWARD AND SEGMENTAL QB LEARNING (THREE EM ITERATIONS, $\mathcal{K} = 8$)

total number of adaptation tokens	on-line adaptation methods		
	fb	seg	ncor
0	51.10	51.10	51.10
5	59.84	59.56	54.31
10	63.28	63.08	58.48
15	69.70	69.26	64.01
20	73.51	73.63	67.90
26	78.84	78.84	77.19
52	84.53	84.85	82.76
78	87.78	87.86	85.73
104	88.38	89.02	87.25
130	89.78	89.94	88.58
156	90.94	90.98	89.70
182	91.18	90.94	89.50
208	91.10	91.26	90.14

VI. DISCUSSION AND CONCLUSION

In this paper, we extend our previously proposed on-line quasi-Bayes adaptive learning framework to handle the correlated CDHMM parameters in which all mean vectors are assumed to be correlated and have a joint Gaussian prior distribution. A successive approximation algorithm is proposed to implement the correlated mean vectors' updating. 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 top $\mathcal{K}$ prediction, the effects of different initial hyperparameters estimations, the effects of different number of EM iterations, and finally the comparison between forward-backward and segmental QB learning. We have found the following.

- The proposed QB and its successive approximation learning algorithm is capable of enhancing the efficiency and the effectiveness of the Bayes learning by taking into account the correlation information between different models as well as having a good asymptotic convergence behavior.
- A good definition of correlation neighborhood and an appropriate choice of the size of the neighborhood is a key for improving the efficacy of on-line adaptation.
- A good initial prior distribution is crucial for improving the efficacy of on-line adaptation. Specifically, in method

of moment estimate, more samples will be helpful for initial correlation coefficients estimation.

- The segmental QB learning with one EM iteration for each OLA step is a good engineering comprise between computational complexity and the performance degradation.

We are also working in the following areas:

- development of techniques to make the current QB learning framework work equally well under unsupervised mode;
- examining the on-line stochastic acoustic normalization techniques and their combinations with the on-line model adaptation;
- formulation and development of the appropriate mathematical tools for a good intrinsic structural model of speech in acoustic, phonetic, and linguistic aspects, which is believed to be crucial for efficient adaptation;
- investigation of new robust decision strategies and their combinations with the on-line model adaptation to further enhance the on-line recognition performance.

Our ultimate goal will be to easily adapt a set of general models to new task, new speaker and new environment.

ACKNOWLEDGMENT

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

REFERENCES

[1] S. M. Ahadi and P. C. Woodland, "Rapid speaker adaptation using model prediction," in *Proc. ICASSP-95*, Detroit, MI, May 1995, pp. I-684–I-687.

[2] S. Cox, "Predictive speaker adaptation in speech recognition," *Comput. Speech Lang.*, vol. 9, pp. 1–17, 1995.

[3] M. H. DeGroot, *Optimal Statistical Decisions*. New York: McGraw-Hill, 1970.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. B*, vol. 39, pp. 1–38, 1977.

[5] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 357–366, 1995.

[6] V. V. Digalakis and L. G. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 294–300, 1996.

[7] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[8] S. Furui, "A training procedure for isolated word recognition systems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 129–136, 1980.

[9] S. Furui, "Recent advances in robust speech recognition," in *Proc. ESCA-NATO Tutorial and Res. Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, Apr. 1997, pp. 11–20.

[10] 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.

[11] H. Hattori and S. Sagayama, "Vector field smoothing principle for speaker adaptation," in *Proc. ICSLP-92*, pp. 381–384.

[12] 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, pp. 334–345, 1995.

[13] _____, "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.

[14] 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 Audio Processing*, vol. 5, pp. 161–172, 1997.

[15] J. Ishii, M. Tonomura, and S. Matsunaga, "Speaker adaptation using tree structured shared-state HMM's," in *Proc. ICSLP-96*, Philadelphia, PA.

[16] 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, pp. 307–309, Mar. 1986.

[17] B.-H. Juang, "Speech recognition in adverse environments," *Comput. Speech Lang.*, vol. 5, pp. 275–294, 1991.

[18] 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, pp. 530–535, 1984.

[19] 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.

[20] 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, 1991.

[21] C.-H. Lee, F.-K. Soong, and K.-K. Paliwal, Eds., *Automatic Speech and Speaker Recognition: Advanced Topics*. Boston, MA: Kluwer, 1996.

[22] C.-H. Lee, "On feature and model compensation approach to robust speech recognition," in *Proc. ESCA-NATO Tutorial and Res. Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, Apr. 1997, pp. 45–54.

[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] K. Ohkura, M. Sugiyama, and S. Sagayama, "Speaker adaptation based on transfer vector field smoothing with continuous mixture density HMM's," in *Proc. ICSLP-92*, pp. 369–372.

[25] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, 1989.

[26] L. R. Rabiner, B.-H. Juang, and C.-H. Lee, "An overview of automatic speech recognition," in *Automatic Speech and Speaker Recognition: Advanced Topics*, C.-H. Lee, F.-K. Soong, and K.-K. Paliwal, Eds. Boston, MA: Kluwer, 1996, pp. 1–30.

[27] O. Ronen and M. Ostendorf, "A dependence tree model of phone correlation," in *Proc. ICASSP-96*, Atlanta, GA, pp. 873–876.

[28] W. A. Rozzi and R. M. Stern, "Speaker adaptation in continuous speech recognition via estimation of correlated mean vectors," in *Proc. ICASSP-91*, Toronto, Ont., Canada, May 1991, pp. 865–868.

[29] A. Sankar and C.-H. Lee, "A maximum likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 190–202, 1996.

[30] B. M. Shahshahani, "A Markov random field approach to Bayesian speaker adaptation," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 183–191, 1997.

[31] K. Shinoda and T. Watanabe, "Speaker adaptation with autonomous model complexity control by MDL principle," in *Proc. ICASSP-96*, Atlanta, GA, May 1996, pp. 717–720.

[32] R. M. Stern and M. J. Lasry, "Dynamic speaker adaptation for feature-based isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 751–763, 1987.

[33] 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.

[34] 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.

[35] M. Tonomura, T. Kosaka, S. Matsunaga, and A. Monden, "Speaker adaptation fitting training data size and contents," in *Proc. EUROSPEECH-95*, Madrid, Spain, pp. 1147–1150.

[36] G. Zavaliagkos, R. Schwartz, and J. Makhoul, "Batch, incremental and instantaneous adaptation techniques for speech recognition," in *Proc. ICASSP-95*, Detroit, MI, pp. I-676–679.

[37] G. Zavaliagkos, R. Schwartz, J. McDonough, and J. Makhoul, "Adaptation algorithms for large scale HMM recognizers," in *Proc. EUROSPEECH-95*, Madrid, Spain, pp. 1131–1134.

[38] Y.-X. Zhao, "An acoustic-phonetic-based speaker adaptation technique for improving speaker-independent continuous speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 380–394, 1994.

Qiang Huo (M'95) received the B.Eng. degree from University of Science and Technology of China (USTC), Hefei, in 1987, the M.Eng. degree from Zhejiang University, Hangzhou, 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, speech and speaker recognition. From 1991 to 1994, he was with the Department of Computer Science, University of Hong Kong (HKU), where he completed his Ph.D. dissertation on speech recognition under a joint Ph.D. training program between HKU and USTC. From April 1995 to December 1997, he was with ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan, where he engaged in research in speech recognition. He joined the Department of Computer Science and Information Systems, HKU, again in January 1998 as an Assistant Professor. 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 the University of Washington, Seattle, in 1981, all in electrical engineering.

In 1981, he joined Verbex Corporation, Bedford, MA, and was involved in research work on connected word recognition. In 1984, he became affiliated with Digital Sound Corporation, Santa Barbara, CA, where he engaged in research in speech coding, speech recognition and signal processing for the development of the DSC-2000 Voice Server. Since 1986, he has been with Bell Laboratories, Murray Hill, NJ, where he is now a Distinguished Member of 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 recent edited book entitled *Automatic Speech and Speaker Recognition: Advanced Topics* (Boston, MA: Kluwer, 1996).

Dr. Lee served as an associate editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and TRANSACTIONS ON SPEECH AND AUDIO PROCESSING from 1991 to 1995. He was a member of the ARPA Spoken Language Coordination Committee between 1991 and 1995. He has also been 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 and 1996 SPS Best Paper Award. He currently serves as the Chairman of the SPS Speech Technical Committee.