Online Hierarchical Transformation of Hidden Markov Models for Speech Recognition

Jen-Tzung Chien, *Member, IEEE*

***Abstract*— This paper proposes a novel framework of online hierarchical transformation of hidden Markov model (HMM) parameters for adaptive speech recognition. Our goal is to incrementally transform (or adapt) all the HMM parameters to a new acoustical environment even though most of HMM units are unseen in observed adaptation data. We establish a hierarchical tree of HMM units and apply the tree to dynamically search the transformation parameters for individual HMM mixture components. In this paper, the transformation framework is formulated according to the approximate Bayesian estimate, which the prior statistics and the transformation parameters can be jointly and incrementally refreshed after each consecutive adaptation data is presented. Using this formulation, only the refreshed prior statistics and the current block of data are needed for online transformation. In a series of speaker adaptation experiments on the recognition of 408 Mandarin syllables, we examine the effects on constructing various types of hierarchical trees. The efficiency and effectiveness of proposed method on incremental adaptation of overall HMM units are also confirmed. Besides, we demonstrate the superiority of proposed online transformation to Huo's on-line adaptation [16] for a wide range of adaptation data.**

***Index Terms*—Approximate Bayesian estimate, EM algorithm, hidden Markov models, online hierarchical transformation, speaker adaptation, speech recognition.**

I. Introduction

It IS generally agreed that an inevitable trend for practical speech recognition system should be developed toward the goals of recognizing speech uttered by a large population of speakers and without limiting the vocabulary size. However, the desirable performance of such large vocabulary and speaker independent (SI) speech recognition system is usually difficult to achieve because of the complex variabilities arising from phonetics and speakers. A fascinating approach to alleviate the difficulties is to design a robust (or adaptive) algorithm to cope with the possible variabilities between testing speech and existing SI reference models (or seed models). For example, a speaker adaptation technique is feasible to improve the SI speech recognizer for a specific speaker by using his adaptation (or enrollment, calibration) utterances. By adapting the SI model parameters to a new speaker (along with his operating environments including transducer, telephone channel and surrounding noise), the mismatch in pattern recognition can be reduced and good recognition performance can be obtained. Moreover, it would be further preferred that the adaptation scheme has the capability to adapt an entire set of large-vocabulary model parameters by using limited adaptation data, of which some sounds are unheard in adaptation data. If we can develop such adaptive techniques, the difficulties in SI and large-vocabulary speech recognition systems may be partially resolved.

Generally, adaptive techniques are employed in three strategies [25], [42]: 1) *batch adaptation*, 2) *self adaptation*, and 3) *online adaptation*. Batch adaptation is an offline adaptation where the models are adapted by using batch data, i.e., adaptation is performed after all adaptation utterances are collected. In recent years, much research has been exploited for the application of batch adaptation [14], [26]. Self adaptation (or referred as *instantaneous adaptation*) executes the adaptation on testing data itself at runtime [6], [43]. It is able to trace the changing conditions during the recognition. However, owing to the insufficient testing observations, the adaptation performance is constrained and therefore the recognition improvement is limited. Besides, online adaptation (or referred to as *incremental adaptation* or *sequential adaptation*) is capable of balancing the tradeoff between batch adaptation and self adaptation. It aims at incrementally tracing the changing variations only when a block of adaptation/training/testing data is enrolled. This block of data is then thrown away after the adaptation is completed. As a consequence, the merit of online adaptation is to continuously update the model parameters without waiting long history of batch data. The computational load and the memory requirement can be efficiently reduced. Its flexible property has been attracting increasing number of studies focused on this issue [9], [15]–[17], [27], [38]. This technique is not only practicable for incremental training of model parameters but also online speaker adaptation. In this paper, we concentrate our efforts on presenting a novel framework of online adaptation.

In the literature, two main categories of adaptation algorithms have been proposed and widely applied for hidden Markov model (HMM) based speech recognition [4], [5]. One is the *transformation-based adaptation*, where the clusters of HMM parameters are individually transformed (or adapted) by means of some transformation functions. The frameworks of maximum likelihood linear regression (MLLR) [26], maximum likelihood stochastic matching (SM) [32] and constrained transformation [10] were classified into this category. In their works, the transformation parameters were derived via the maximum likelihood (ML) estimate. In fact, we can also utilize

Manuscript received April 16, 1998; revised February 1, 1999. This work was supported in part by the National Science Council, Taiwan, R.O.C. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Richard C. Rose.

The author is with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 70101, Taiwan, R.O.C. (e-mail: jtchien@mail.ncku.edu.tw).

Publisher Item Identifier S 1063-6676(99)07985-7.

the maximum *a posteriori* (MAP) criterion to incorporate prior information into transformation [3], [6]. On the other hand, a series of studies on the *MAP adaptation* (or learning) of HMM parameters are grouped into the second category. By serving the SI HMM parameters as prior statistics, the HMM parameters including mixture weight, mean vector, and covariance matrix were adapted accordingly based on the MAP theory [14], [18], [23]. If infinite training data are provided, MAP estimate is proved to asymptotically approach to ML estimate. In general, when the adaptation data is limited, the transformation-based adaptation can efficiently transform all the HMM parameters according to cluster-dependent transformation functions. Conversely, when the adaptation data is abundant, the MAP adaptation of HMM parameters can effectively merge the adaptation tokens with the corresponding SI HMM parameters. Our previous studies reported that the hybrid algorithm of transformation-based adaptation and MAP adaptation attained a better recognition performance than transformation or adaptation alone for any practical amount of adaptation data [4], [5]. Moreover, in transformation-based adaptation, the construction of hierarchical tree of HMM parameters provides an effective approach to dynamically capture the goodness of transformation parameters and elevate the recognition performance for various data amounts. The related works include the autonomous model complexity control (AMCC) algorithm [33], the transformation smoothing method [13], and the structural MAP (SMAP) approach [34], [35].

From the explanation above, we are motivated to propose a novel online hierarchical transformation algorithm for robust HMM-based speech recognition. The underlying theoretical foundation is based on the approximate Bayesian (quasi-Bayes, or QB) estimate described by Huo and Lee [16], [17]. According to this estimate, the unknown parameters are estimated by maximizing an approximate posterior distribution, which is a product of a likelihood function of current block data and a prior density given the updated parameter statistics (or hyperparameters). The hyperparameters can be obtained from previous observed data. By specifying the prior density as a conjugate prior, we may generate a *reproducible prior/posterior pair*. Then, a recursive MAP estimate can be derived and applied for incrementally executing adaptation and updating hyperparameters. In the work of Huo and Lee [16], an approximate recursive Bayes learning of continuous-density HMM (CDHMM) parameters was formulated and successfully applied for speaker adaptation using the 26-letter English alphabet vocabulary. Their experiments relied on the speaker providing at least one example of each vocabulary in adaptation data. When the enrolled adaptation data are incomplete, parts of HMM parameters are never adapted and therefore the recognition performance is restricted. Further, they extended the quasi-Bayes estimate to cope with the correlated CDHMM parameters in which all HMM mean vectors are correlated with a joint prior distribution [17]. This algorithm provided a scheme to adapt the unseen HMM units in adaptation data by considering the correlation between different mean vectors. In this paper, we present a novel transformation-based online adaptation approach, where the overall HMM parameters are incrementally adapted through a set of transformation functions. Instead of [16], [17], the QB estimate is applied for estimating the transformation parameters for adapting HMM parameters. For the sake of automatically capturing the goodness of transformation parameters, we build a hierarchical tree of HMM parameters such that each HMM unit is capable of searching its most likely transformation parameters from leaf node to root node. For each HMM unit, we extract the node containing the adaptation tokens and use its parameters for online transformation. This algorithm is also referred as *online hierarchical transformation*. From the experimental results, we find that the proposed method is feasible to speaker adaptation either in incremental mode or batch mode for various numbers of adaptation data.

This paper is organized as follows. In next section, the theoretical formulation of online transformation is addressed. The estimation of initial hyperparameters and the segmental QB approach are also included. Section III describes the establishment of hierarchical tree and the use of structural transformation in proposed online transformation. The experimental setup and databases are also mentioned in this section. Following, a series of comparative experiments on speaker adaptation are carried out in Section IV. Some observations and discussions are reported. Finally, the conclusion is given in Section V.

II. Online Transformation

Among the previous studies, the on-line parameter estimation schemes based on incremental expectation-maximization (EM) algorithm [8] were derived by using stochastic approximations maximizing the Kullback–Leibler information measure [22], [41]. The incremental generalized EM algorithm was also used for supervised learning of hierarchical mixture models [19] and gradient-based training of HMM's [2]. Recently, Digalakis [9] employed the incremental EM algorithms in estimation of affine transformation parameters. In this paper, an attractive approach based on QB estimate [16], [17], [36] is considered as the theoretical basis for online estimation of transformation parameters. The estimated parameters are used for transforming the HMM parameters to a new environment.

A. General Formulation

Consider an N-state CDHMM with state parameters composed of K mixture components, $\lambda = \{\lambda_i\} = \{\omega_{ik}, \mu_{ik}, r_{ik}\}$, $i = 1, \cdots, N$, $k = 1, \cdots, K$, the state observation probability density function (pdf) of sample $\mathbf{x}$ is assumed to be a mixture of multivariate Gaussian distributions

$$p(\mathbf{x} \mid \lambda_i) = \sum_{k=1}^{K} \omega_{ik} N(\mathbf{x} \mid \mu_{ik}, r_{ik}) \tag{1}$$

where ω_{ik} is the mixture gain subject to the constraint $\sum_{k=1}^{K} \omega_{ik} = 1$, μ_{ik} is the d-dimensional mean vector, r_{ik} is the $d \times d$ precision (or inverse covariance) matrix, $r_{ik} = \sum_{ik}^{-1}$, and $N(\mathbf{x} \mid \mu_{ik}, r_{ik})$ is the Gaussian distribution denoted by

$$N(\mathbf{x} \mid \mu_{ik}, r_{ik}) \propto |r_{ik}|^{1/2} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu_{ik})^T r_{ik} (\mathbf{x} - \mu_{ik})\right]. \tag{2}$$

In online transformation, the HMM parameters λ are transformed according to a transformation function $G_\eta(\cdot)$. Let $\chi^n = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n\}$ be n i.i.d. and successively observed samples/blocks, which are served as the adaptation data to estimate the transformation parameters η. The *a posteriori* density of η satisfies the following recursive relation [11], [37]:

$$p(\eta \mid \chi^n) = \frac{p(\mathbf{X}_n \mid \eta) \cdot p(\eta \mid \chi^{n-1})}{\int p(\mathbf{X}_n \mid \eta) \cdot p(\eta \mid \chi^{n-1})\, d\eta}. \tag{3}$$

Repeating use of (3) can produce a recursive formula of posterior density. This provides a motivation for the recursive Bayesian estimate. However, because of the missing data problem in CDHMM framework, the recursive Bayesian estimate using (3) will greatly raise the computational complexity [16]. To alleviate the computational difficulties, an approximate posterior density is adopted. Hence, the approximate MAP (or QB) estimate of $\eta^{(n)}$ after observing the current sample $\mathbf{X}_n$ is developed by

$$\begin{aligned}\eta^{(n)} &= \arg\max_{\eta} p(\eta \mid \chi^n) = \arg\max_{\eta} p(\mathbf{X}_n \mid \eta) \cdot p(\eta \mid \chi^{n-1}) \\ &\cong \arg\max_{\eta} p(\mathbf{X}_n \mid \eta) \cdot g\big(\eta \,\big|\, \varphi^{(n-1)}\big).\end{aligned} \tag{4}$$

Herein, the true posterior density of previous accumulated data $p(\eta \mid \chi^{n-1})$ is approximated by the closest tractable prior density $g(\eta \mid \varphi^{(n-1)})$ where $\varphi^{(n-1)}$ denotes the updated hyperparameters after observing previous χ^{n-1}. The hyperparameters $\varphi^{(n-1)}$ used in this paper will be defined in Section II-B. Based on the QB estimate with initial hyperparameters $\varphi^{(0)}$, we can estimate the transformation parameters $\eta^{(1)}$ by applying $\mathbf{X}_1$ in (4). Then, the hyperparameters $\varphi^{(1)}$ are updated and stored for the estimation of next parameters $\eta^{(2)}$. Accordingly, a recursive QB formulation for parameter sequence $\eta^{(1)}, \eta^{(2)}, \cdots, \eta^{(n)}$ can be established.

Because the QB estimate in (4) is an incomplete data problem, we use the EM algorithm to iteratively improve the approximate posterior likelihood of current estimate $\eta^{(n)}$ and derive the new estimate $\hat{\eta}^{(n)}$ in an optimal manner [8]. By applying the EM algorithm, the QB estimate is completed by iteratively performing the following two steps.

1) E-Step: Calculate the auxiliary function

$$\begin{aligned}R\big(\hat{\eta}^{(n)} \,\big|\, \eta^{(n)}\big) = E\big\{&\log p\big(\mathbf{X}_n, \mathbf{s}_n, \mathbf{l}_n \mid \hat{\eta}^{(n)}\big) \\ &+ \log g\big(\hat{\eta}^{(n)} \,\big|\, \varphi^{(n-1)}\big) \big| \mathbf{X}_n, \eta^{(n)}\big\}\end{aligned} \tag{5}$$

where $\mathbf{s}_n = \{s_t^{(n)}\}$ is the state sequence, $\mathbf{l}_n = \{l_t^{(n)}\}$ is the mixture component sequence, and $(\mathbf{X}_n, \mathbf{s}_n, \mathbf{l}_n)$ is our choice of complete data.

2) M-Step: Find the new estimate

$$\hat{\eta}^{(n)} = \arg\max_{\hat{\eta}^{(n)}} R\big(\hat{\eta}^{(n)} \,\big|\, \eta^{(n)}\big). \tag{6}$$

The iterative EM steps guarantee that the approximate posterior likelihood never decreases. In [3], [6], and [16], a forgetting (or tuning) factor was incorporated to tune the importance of likelihood function and prior density in MAP estimation. This issue is not considered herein. In the next section, we specify the form of transformation function and the distribution family of prior density. A set of QB formulas can be derived for online transformation.

B. Derivation of Online Transformation Parameters

Several kinds of transformation functions have been investigated for compensating the mismatch between testing speech and model parameters, including bias transformation [30], affine transformation [10], [26], nonlinear transformation [1], and vector Taylor series transformation [28], etc. We also addressed a Bayesian affine transformation approach to batch adaptation [3], [6]. In this study, the transformation of CDHMM parameters $\lambda = \{\omega_{ik}, \mu_{ik}, r_{ik}\}$ is constrained in a form of

$$\hat{\lambda} = G_{\eta^{(n)}}(\lambda) = \big\{\omega_{ik}, \mu_{ik} + \mu_c^{(n)}, \theta_c^{(n)} r_{ik}\big\} \tag{7}$$

where mean vector μ_{ik} is transformed by a bias vector $\mu_c^{(n)}$, precision matrix r_{ik} is transformed by a $d \times d$ scaling matrix $\theta_c^{(n)}$ and mixture gain ω_{ik} is not adapted. Herein, the transformation function $G_{\eta^{(n)}}(\cdot)$, $\eta^{(n)} = \{\eta_c^{(n)}\} = \{\mu_c^{(n)}, \theta_c^{(n)}\}$, $c = 1, \cdots, C$, having C clusters is used. The HMM parameters $\lambda_{ik} = (\mu_{ik}, r_{ik})$ are assumed to be labeled by the cth cluster membership Ω_c.

Furthermore, another important issue of QB estimate is the choice of prior density family. As indicated in [14] and [23], the prior density in *conjugate family* is a good candidate due to the reason of mathematical attractiveness. Herein, we also obey this rule and define the joint prior density of transformation parameters $\eta_c^{(n)} = (\mu_c^{(n)}, \theta_c^{(n)})$ of membership Ω_c to be a normal-Wishart density of the form [7]

$$\begin{aligned}g\big(\mu_c^{(n)}, \theta_c^{(n)}\big) &= g\big(\eta_c^{(n)} \,\big|\, \varphi_c^{(n-1)}\big) \propto \big|\theta_c^{(n)}\big|^{(\alpha_c^{(n-1)} - d)/2} \\ &\quad \times \exp\Big[-\frac{1}{2}\big(\mu_c^{(n)} - m_c^{(n-1)}\big)^T \theta_c^{(n)} \tau_c^{(n-1)} \\ &\quad \times \big(\mu_c^{(n)} - m_c^{(n-1)}\big)\Big] \\ &\quad \times \exp\Big[-\frac{1}{2}\operatorname{tr}\big(u_c^{(n-1)} \theta_c^{(n)}\big)\Big]\end{aligned} \tag{8}$$

where $\varphi_c^{(n-1)} = (\tau_c^{(n-1)}, m_c^{(n-1)}, \alpha_c^{(n-1)}, u_c^{(n-1)})$ are the hyperparameters of prior density such that $\alpha_c^{(n-1)} > d - 1$, $m_c^{(n-1)}$ is a d-dimensional vector, $\tau_c^{(n-1)}$ and $u_c^{(n-1)}$ are $d \times d$ symmetric positive definite matrices. The notation $\operatorname{tr}(\cdot)$ represents the trace of a matrix. The hyperparameters $\varphi^{(n-1)} = \{\varphi_c^{(n-1)}\}$ can be obtained from the previous enrolled data χ^{n-1}. Under this definition, the posterior density of complete data [i.e., exponential of auxiliary function in (5)] multiplied by a normalization term K, $K \cdot \exp\{R(\hat{\eta}_c^{(n)} \mid \eta_c^{(n)})\}$, can be also expressed in a form of normal-Wishart density $g(\hat{\eta}_c^{(n)} \mid \hat{\varphi}_c)$ with the new hyperparameters $\hat{\varphi}_c = (\hat{\tau}_c, \hat{m}_c, \hat{\alpha}_c, \hat{u}_c)$ given as follows (see Appendix for derivation):

$$\hat{\tau}_c = \tau_c^{(n-1)} + \sum_{i,k \in \Omega_c} c_{ik} r_{ik}, \tag{9}$$

$$\begin{aligned}\hat{m}_c &= \left(\tau_c^{(n-1)} + \sum_{i,k \in \Omega_c} c_{ik} r_{ik}\right)^{-1} \\ &\quad \cdot \left(\tau_c^{(n-1)} m_c^{(n-1)} + \sum_{i,k \in \Omega_c} c_{ik} r_{ik} \bar{\mathbf{b}}_c\right)\end{aligned} \tag{10}$$

$$\hat{\alpha}_c = \alpha_c^{(n-1)} + \sum_{i,k\in\Omega_c} c_{ik}, \tag{11}$$

$$\begin{aligned}\hat{u}_c = u_c^{(n-1)} &+ \sum_{i,k\in\Omega_c} S_{ik}r_{ik} + \left(\tau_c^{(n-1)} + \sum_{i,k\in\Omega_c} c_{ik}r_{ik}\right)^{-1} \\ &\cdot \left(\tau_c^{(n-1)} \sum_{i,k\in\Omega_c} c_{ik}r_{ik}\right) \cdot \left(\bar{\mathbf{b}}_c - m_c^{(n-1)}\right) \\ &\cdot \left(\bar{\mathbf{b}}_c - m_c^{(n-1)}\right)^T\end{aligned} \tag{12}$$

where $\xi_t(i,k) = \Pr(s_t^{(n)} = i, l_t^{(n)} = k \mid \mathbf{X}_n, \eta_c^{(n)})$ is the posterior probability of being in state i and mixture component k given that the current parameters $\eta_c^{(n)}$ generate $\mathbf{X}_n = \{\mathbf{x}_t^{(n)}\}$ and

$$c_{ik} = \sum_t \xi_t(i,k), \tag{13}$$

$$\bar{\mathbf{b}}_c = \sum_t \sum_{i,k\in\Omega_c} \xi_t(i,k)\left(\mathbf{x}_t^{(n)} - \mu_{ik}\right) \Bigg/ \sum_t \sum_{i,k\in\Omega_c} \xi_t(i,k), \tag{14}$$

$$S_{ik} = \sum_t \xi_t(i,k)\left(\mathbf{x}_t^{(n)} - \mu_{ik} - \bar{\mathbf{b}}_c\right)\left(\mathbf{x}_t^{(n)} - \mu_{ik} - \bar{\mathbf{b}}_c\right)^T. \tag{15}$$

The E-step of EM algorithm is therefore completed. Notably, the use of matrix in hyperparameter $\tau_c^{(n-1)}$ is to generalize the estimation of full hyperparameters $\hat{m}_c$ and $\hat{u}_c$. In the M-step, we maximize $g(\hat{\eta}_c^{(n)} \mid \hat{\varphi}_c)$ with respect to $\hat{\eta}_c^{(n)}$ and derive the new estimate of transformation parameters $\hat{\eta}_c^{(n)} = (\hat{u}_c^{(n)}, \hat{\theta}_c^{(n)})$ as follows:

$$\hat{\mu}_c^{(n)} = \hat{m}_c, \tag{16}$$

$$\hat{\theta}_c^{(n)^{-1}} = (\hat{\alpha}_c - d)^{-1}\hat{u}_c. \tag{17}$$

By iteratively performing E-step and M-step for several times, we finally obtain the transformation parameters $\hat{\eta}_c^{(n)}$. Using $\hat{\eta}^{(n)} = \{\hat{\eta}_c^{(n)}\} = \{\hat{u}_c^{(n)}, \hat{\theta}_c^{(n)}\}$, the HMM parameters are transformed according to (7). Notably, this set of formulas can be easily extended to the case of diagonal matrices for the CDHMM precision matrix r_{ik} and the transformation matrix $\theta_c^{(n)}$. In this case, by setting the prior transformation pdf for each diagonal component to be a normal-gamma density [7], we can similarly derive a new set of formulas. This derivation is neglected in this study.

After the transformation, the hyperparameters are refreshed by

$$\begin{aligned}\varphi^{(n)} &= \{\varphi_c^{(n)}\} = \{\tau_c^{(n)}, m_c^{(n)}, \alpha_c^{(n)}, u_c^{(n)}\} = \{\hat{\varphi}_c\} \\ &= \{\hat{\tau}_c, \hat{m}_c, \hat{\alpha}_c, \hat{u}_c\}.\end{aligned} \tag{18}$$

These hyperparameters $\varphi^{(n)}$ are then kept in memory and served as the new hyperparameters for online estimation of next transformation parameters $\eta^{(n+1)} = \{\eta_c^{(n+1)}\}$ when consecutive data $\mathbf{X}_{n+1}$ are collected. As shown in above derivation, the advantage of proposed method is mainly focused on the generation of *reproducible prior/posterior pair* in EM algorithm so that the transformation parameters and the hyperparameters can be efficiently and recursively computed for online transformation. To initialize the online transformation, we have to estimate the initial hyperparameters $\varphi^{(0)}$ when the first block of data $\mathbf{X}_1$ is enrolled for model transformation.

C. Estimation of Initial Hyperparameters

Basically, the initial hyperparameters should provide the sufficient knowledge of unknown parameters such that the reliable parameters can be incrementally produced. In the special application of speaker adaptation, the initial hyperparameters should reflect the physical meaning of transformation factors from a large population of speakers. The goodness of estimated transformation parameters could be accordingly verified. In this study, the initial hyperparameters are estimated in such an *empirical Bayes* sense [31]. We try to extract the meaningful hyperparameters $\varphi^{(0)} = \{\tau_c^{(0)}, m_c^{(0)}, \alpha_c^{(0)}, \mu_c^{(0)}\}$ from SI training data [16].

Let $\chi_1, \cdots, \chi_Q$ denote the training data of speakers $q = 1, \cdots, Q$ with $\chi_q = \{\mathbf{x}_{q,t}\}$ and $\tilde{\mathbf{b}}_{cq}$ represents the averaged bias vector between the training utterances of the speaker q and the SI model parameters related to HMM cluster c. The initial hyperparameters adopted in the following experiments are estimated by

$$\begin{aligned}m_c^{(0)} &= E\{\tilde{\mathbf{b}}_{cq}\} \\ &= \frac{1}{Q}\sum_{q=1}^{Q} \frac{\sum_t \sum_{i,k\in\Omega_c} \xi_{q,t}^{(\mathrm{SI})}(i,k)(\mathbf{x}_{q,t} - \mu_{ik})}{\sum_t \sum_{i,k\in\Omega_c} \xi_{q,t}^{(\mathrm{SI})}(i,k)}\end{aligned} \tag{19}$$

$$\tau_c^{(0)} = \sum_{i,k\in\Omega_c} c_{ik}^{(\mathrm{SI})} r_{ik} \Bigg/ \sum_{i,k\in\Omega_c} c_{ik}^{(\mathrm{SI})} \tag{20}$$

$$\alpha_c^{(0)} = d + 1 \tag{21}$$

and (22), shown at the bottom of the page, where the posterior probabilities $\xi_{q,t}^{(\mathrm{SI})}(i,k) = \Pr(s_{q,t} = i, l_{q,t} = k \mid \mathbf{x}_{q,t}, \lambda)$ are determined by the last iteration of HMM training of SI speech data and

$$c_{ik}^{(\mathrm{SI})} = \sum_q \sum_t \xi_{q,t}^{(\mathrm{SI})}(i,k). \tag{23}$$

Note that $m_c^{(0)}$ and $u_c^{(0)}$ are, respectively, obtained by taking the mean and the variance of $\{\tilde{\mathbf{b}}_{cq}\}$, $c = 1, \ldots, C$, $q = 1, \ldots, Q$, with respect to the speaker index q.

$$u_c^{(0)} = \mathrm{Var}\{\tilde{\mathbf{b}}_{cq}\} = \frac{1}{Q}\sum_{q=1}^{Q} \frac{\sum_t \sum_{i,k\in\Omega_c} \xi_{q,t}^{(\mathrm{SI})}(i,k)\left(\mathbf{x}_{q,t} - \mu_{ik} - m_c^{(0)}\right)\left(\mathbf{x}_{q,t} - \mu_{ik} - m_c^{(0)}\right)^T}{\sum_t \sum_{i,k\in\Omega_c} \xi_{q,t}^{(\mathrm{SI})}(i,k)} \tag{22}$$

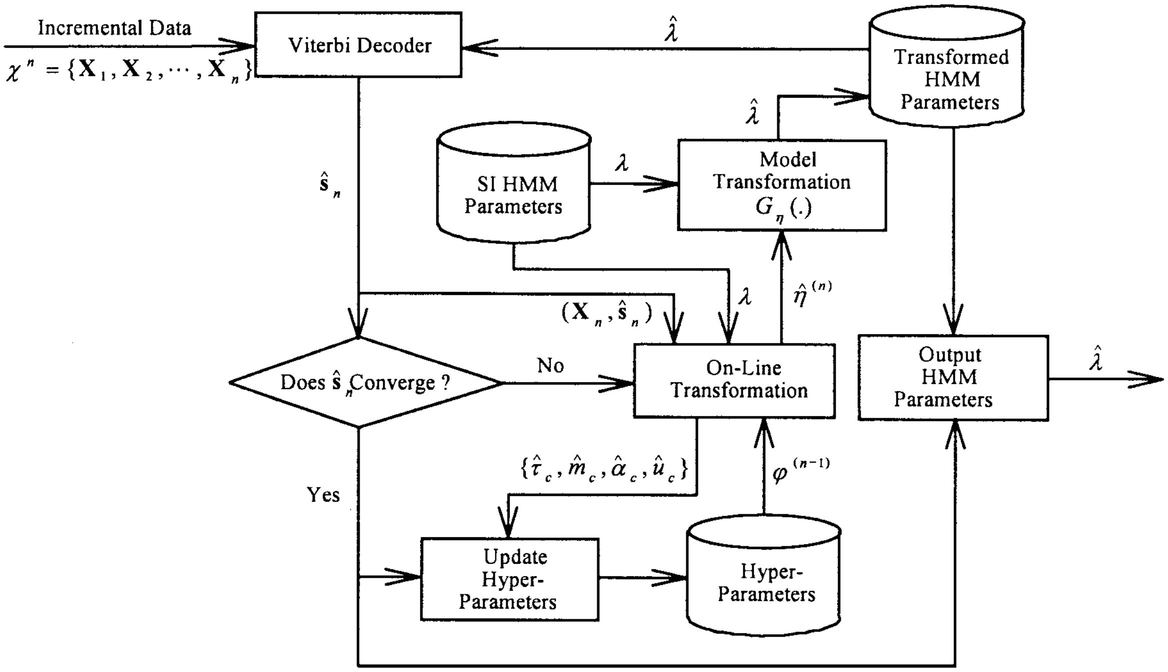

Fig. 1. Block diagram of online transformation of HMM parameters based on segmental QB estimate.

D. Segmental QB Estimate

In general, the formulation in Section II-B is also termed as the *forward-backward QB estimate*. The posterior density $\xi_t(i,k) = \Pr(s_t^{(n)} = i, l_t^{(n)} = k \mid \mathbf{X}_n, \eta_c^{(n)})$ can be efficiently computed via the forward-backward algorithm [20]. Alternatively, the forward-backward QB estimate can be extended to the *segmental QB estimate* in which the joint posterior likelihood of transformation parameters η and state sequence $\mathbf{s}_n$, $p(\eta, \mathbf{s}_n \mid \chi^n)$, is maximized. The estimation procedure becomes

$$\begin{aligned}\eta^{(n)} &= \arg\max_{\eta}\max_{\mathbf{s}_n} p(\eta, \mathbf{s}_n \mid \chi^n)\\ &\cong \arg\max_{\eta}\max_{\mathbf{s}_n} p(\mathbf{X}_n, \mathbf{s}_n \mid \eta)\cdot g\big(\eta \,\big|\, \varphi^{(n-1)}\big) \qquad (24)\end{aligned}$$

which can be further divided into the following two equations:

$$\hat{\mathbf{s}}_n = \arg\max_{\mathbf{s}_n} p(\mathbf{X}_n, \mathbf{s}_n \mid \eta), \qquad (25)$$

$$\hat{\eta}^{(n)} = \arg\max_{\eta} p(\mathbf{X}_\eta, \hat{\mathbf{s}}_n \mid \eta)\cdot g\big(\eta \,\big|\, \varphi^{(n-1)}\big). \qquad (26)$$

According to (25)–(26), the most likely state sequence $\hat{\mathbf{s}}_n$ is first decoded by using the Viterbi algorithm [40]. Given $\hat{\mathbf{s}}_n$, the new QB estimate $\hat{\eta}^{(n)}$ is obtained by (26). Similar to forward-backward procedure, we can solve (26) by applying the EM algorithm again. Based on EM algorithm, it can be shown that the reestimation equations of (9)–(17) still hold for the segmental QB estimate except that the posterior density $\xi_t(i,k) = \Pr(s_t^{(n)} = i, l_t^{(n)} = k \mid \mathbf{X}_n, \eta_c^{(n)})$ is replaced by

$$\xi_t(i,k) = \delta\big(\hat{s}_t^{(n)} - i\big)\frac{\omega_{ik} N\big(\mathbf{x}_t^{(n)} \,\big|\, \mu_{ik}, r_{ik}, \eta_c^{(n)}\big)}{\sum_{m=1}^{K} \omega_{im} N\big(\mathbf{x}_t^{(n)} \,\big|\, \mu_{im}, r_{im}, \eta_c^{(n)}\big)} \qquad (27)$$

where $\delta(\cdot)$ is the Kronecker delta function. Based on the segmental QB estimate, the online transformation is implemented and illustrated in Fig. 1. In the following section, we address the construction of HMM's tree structure and apply it to autonomously control the goodness of parameters for online hierarchical transformation.

III. Construction of Hierarchical Transformation

In previous studies, the tree-structured clustering technique was applied to build the hierarchical clusters of reference speakers and enabled unsupervised speaker adaptation in SI speech recognition [21]. By selecting the acoustically closest subset to test speaker, quick speaker adaptation was achieved [29]. In contrast, the purpose of hierarchical clustering in proposed method is to establish a tree structure of existing SI HMM units. Based on the HMM clustering labels in various tree levels, we are able to search the most fitted parameters for online transformation of individual HMM units. Before further description of hierarchical transformation, we introduce our experimental setup and databases.

A. Experimental Setup and Databases

The experiments conducted in this paper are aimed at the recognition of Mandarin speech. Mandarin is a syllabic and tonal language. Without considering the tonal information, the overall number of Mandarin syllable is 408. Generally, each Mandarin syllable can be divided into an initial part and a final part. The initial part corresponds to a consonant and the final part corresponds to a vowel. When the syllable only has final part, a null initial exists practically. In this study, we employed the context-dependent subsyllable modeling for constructing the HMM units of Mandarin speech [3]. Cumulatively, there were 93 context-dependent (CD) initials, 38

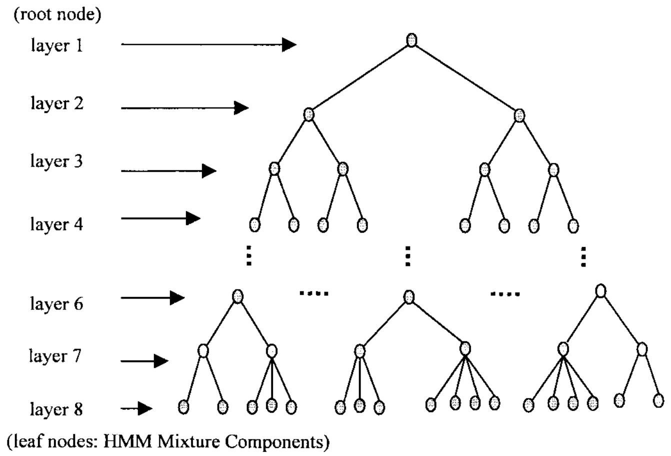

Fig. 2. Hierarchical tree of CDHMM parameters for online transformation.

context-independent (CI) finals and 33 null initials included in the experiments. We arranged the CD initials, CI finals and null initials by three, four, and two HMM states, respectively. Hence, 498 HMM states (279 for CD initials, 152 for CI finals, 66 for null initials, and one for background silence) were setup for covering all phonetic units of 408 Mandarin syllables. Each HMM state has four mixture components at most. In this study, two severely mismatched speech corpora were collected and provided by Telecommunication Laboratories, Chunghwa Telecom, Taiwan. The first one consisted of 5045 phonetically balanced Mandarin words uttered by 51 males and 50 females. It was recorded in an office room via a high-quality microphone. Each Mandarin word contained two to four Mandarin syllables. We applied this database to generate the SI HMM parameters and the initial hyperparameters for online transformation. The speech frame was characterized by a feature vector comprised of 12-order LPC-derived cepstral coefficients, 12-order delta cepstral coefficients, one delta log energy and one delta delta log energy [24]. Besides, the second database consisted of four repetitions of 408 isolated Mandarin syllables spoken by a single female speaker who did not appear in training speakers. This database was collected in a soundproof room with a microphone different from the one used in training database. We used three repetitions for testing and the remaining one for adaptation. Basically, the mismatches between two databases cover changing speakers, microphones and ambient noises.

B. Tree Structure of HMM Parameters

In online transformation, it is crucial to dynamically control the number of transformation parameters such that the recognition accuracy can be improved for limited adaptation data as well as abundant adaptation data. To achieve this goal, a hierarchical structure of HMM parameters should be established prior to performing the adaptation. In this study, the mixture Gaussian densities (or pdf's) of HMM's are clustered by using the K-means algorithm [39] on the basis of binary split. The number of clusters is preset to be 64. After the clustering, a tree structure of HMM parameters with eight layers is accordingly built. Fig. 2 illustrates the hierarchical tree of HMM parameters used in proposed online transformation. Each HMM unit λ_{ik} corresponds to eight nodes for different layers. The root node contains all HMM units. The leaf nodes are occupied by individual HMM units. During the clustering process, the assignment of the mixture component pdf to cluster is based on a predefined distance measure between the mixture component pdf $\lambda_{ik} = (\mu_{ik}, \Sigma_{ik})$ and the pdf (μ_{cb}, Σ_{cb}) drawn from the codebook of pdf's representing the cluster center. Herein, we investigate three distance measures in construction of hierarchical tree. The first one is the weighted Euclidean distance of the form

$$d_{\text{euc}} = (\mu_{ik} - \mu_{cb})^T \Sigma_{cb}^{-1} (\mu_{ik} - \mu_{cb}). \tag{28}$$

It is a popular and efficient distance measure. The second one is the Bhattacharyya distance given by

$$\begin{aligned} d_{\text{Bha}} = {} & \frac{1}{8}(\mu_{ik} - \mu_{cb})^T \left(\frac{\Sigma_{ik} + \Sigma_{cb}}{2} \right)^{-1} (\mu_{ik} - \mu_{cb}) \\ & + \frac{1}{2} \ln \frac{|(\Sigma_{ik} + \Sigma_{cb})/2|}{|\Sigma_{ik}|^{1/2} |\Sigma_{cb}|^{1/2}}. \end{aligned} \tag{29}$$

The Bhattacharyya distance is usually considered as a good measure of separability between two pdf's [12]. In addition, the third distance measure used here is the divergence measure written by

$$\begin{aligned} d_{\text{div}} = {} & \frac{1}{2} \operatorname{tr}\left[(\Sigma_{ik} - \Sigma_{cb}) \left(\Sigma_{cb}^{-1} - \Sigma_{ik}^{-1} \right) \right] \\ & + \frac{1}{2} \operatorname{tr}\left[\left(\Sigma_{ik}^{-1} + \Sigma_{cb}^{-1} \right) (\mu_{ik} - \mu_{cb})(\mu_{ik} - \mu_{cb})^T \right]. \end{aligned} \tag{30}$$

This distance measure corresponds to the total average information for discriminating two pdf's [39]. It can be used to determine pdf's ranking and to evaluate the effectiveness of class discrimination. In our experiments, we compare the performances obtained by using these three distance measures.

C. Hierarchical Transformation

After the hierarchical tree is built, the labels of HMM units in each layer are determined and stored in a lookup table. According to the table, we can calculate the transformation parameters of each tree node by using proposed online transformation paradigm. Theoretically, the HMM units connected to the same node possess similar acoustical behaviors. They can be suitably transformed via the same transformation parameters. Undoubtedly, after the incremental data χ^n are applied to calculate the parameters $\eta^{(n)}$ of each hierarchical node, the nodes in higher layer often accumulate larger amount of adaptation tokens. In case of insufficient adaptation data, part of nodes in lower layer may lack the adaptation tokens. As a result, we usually obtain the transformation parameters for most nodes in higher layer and few nodes in lower layer. In general, the parameters in higher layer are served as *global transformation* and those in lower layer are served as *local transformation*. To reinforce the discriminability of online transformation, the HMM parameters should be transformed as locally as possible. Thus, our aim is to automatically extract the transformation parameters of each HMM unit based on a *bottom-up* search strategy. This strategy captures the transformation factors along the hierarchical path corresponding to each HMM unit. The algorithm of bottom-up search strategy is described below.

Bottom-up search algorithm for on-line transformation parameters

```
1)  for each HMM unit $\lambda_{ik}$
2)    for tree depth from leaf layer to root layer
3)      Extract cluster label of $\lambda_{ik}$ in that depth
4)        if transformation parameters of that label $\eta_c^{(n)}$ exist
5)          Perform on-line transformation $G_{\eta_c^{(n)}}(\lambda_{ik})$
6)            go to step 1
7)          else if hyperparameters of that label $\varphi_c^{(n-1)}$ exist
8)            Perform on-line transformation $G_{\varphi_c^{(n-1)}}(\lambda_{ik})$
9)            go to step 1
10)           end
11)         end
12)   end
```

For each HMM unit λ_{ik}, we search the transformation parameters from leaf layer (i.e., depth $= 8$) to root layer (i.e., depth $= 1$) and perform the following steps. First, the cluster label of λ_{ik} in a layer is extracted. Then, we check if there exist the transformation parameters for this label. If exist, we use the associated parameters $\eta_c^{(n)}$ for on-line transformation, i.e. $G_{\eta_c^{(n)}}(\lambda_{ik})$. Otherwise, we further check if the hyperparameters of this label $\varphi_c^{(n-1)} = (\tau_c^{(n-1)}, m_c^{(n-1)}, \alpha_c^{(n-1)}, u_c^{(n-1)})$ exist. If exist, we transform the mean vector μ_{ik} by adding the bias term $m_c^{(n-1)}$ and the covariance matrix Σ_{ik} by multiplying the scalar term $(\alpha_c^{(n-1)} - d)^{-1}\mu_c^{(n-1)}$ as indicated in (16)–(17). Once the HMM unit λ_{ik} is transformed, we skip to process the next HMM unit. Finally, this algorithm is ended until all the HMM units are transformed.

In online hierarchical transformation, the hyperparameters play an important role because of their utilities in QB parameter estimation and bottom-up transformation algorithm. Herein, we address two issues related to the hyperparameters. First, as mentioned in Section II-B, we have to refresh the hyperparameters following the online transformation. For large-scale HMM framework, the use of incremental data can only refresh a small part of hyperparameters in lower layer. Nevertheless, when the adaptation data appear sequentially, more adaptation tokens and HMM units are collected. The number of refreshed hyperparameters in lower layer can be increased accordingly. Therefore, using proposed method, we can consistently improve the goodness of transformation parameters and their associated hyperparameters for increasing adaptation data. The resulting recognition performance can be improved as well. On the other hand, a practical problem usually happens in estimation of new hyperparameters (9)–(12). As addressed in Section II-C, the initial hyperparameters obtained from training data are served as our prior knowledge of transformation. We drop them out after the first block of data is used for online transformation. Although some hyperparameters are updated following the transformation, most of hyperparameters in lower layer are still empty. This situation results in missing hyperparameters for online transformation of next block of data. When such a situation happens, we capture the hyperparameters of its associated father node. This procedure is then carried out until the existing hyperparameters are captured. Using the approximate hyperparameters, we can calculate the new hyperparameters of (9)–(12) and use them to estimate the transformation parameters in (16)–(17).

IV. Experiments

A series of comparative experiments on incremental speaker adaptation were conducted to demonstrate the merits of proposed method. Readers may refer to Section III-A for detail description of experimental setup and databases. In the experiments, we aim at online adapting the existing SI speech models to a new female speaker by using some adaptation data. Only supervised adaptation was investigated. The adaptation data was sampled from a repetition of 408 Mandarin syllables uttered by this female speaker. The adapted speech models were then applied to recognize her other three repetitions of 408 Mandarin syllables, which is known to be a highly confusable vocabulary. In total, there were 1224 $(3 * 408)$ testing utterances. The average length of a Mandarin syllable including presilence and postsilence was about 0.55 s. The EM iteration number was fixed to be three for all of the experiments. Our recognition system was built on the basis of CDHMM framework. The baseline result using SI speech models (i.e., no adaptation) had a top five recognition rate of 73.8%. Generally, the adaptation performance is sensitive to the presentation and the phonetic content of incremental data especially for small amount of adaptation data. Without loss of generality, we determine the recognition rate of using N adaptation utterances by averaging ten different rounds of recognition task, where each round was accomplished by using N adaptation utterances randomly selected from the adaptation data set. The resulting recognition rate will be more consistent and objective. In this study, we highlight the adaptation performance obtained when most of HMM units are unseen in limited adaptation data. Herein, we plot the relation between occurrence rate of HMM units and number of adaptation data (N) in Fig. 3. In case of $N = 5$, only 6.6% of HMM units was occurred. Even if $N = 100$, the occurrence rate of HMM units was 64.1%. Using proposed hierarchical transformation, all of HMM units can be effectively transformed as demonstrated in the following sections. In the experiments, the CDHMM precision matrix and the transformation matrix $\theta_c^{(n)}$ were set to be diagonal. The following experiments assess the proposed method by the depths and the distance measures used in construction of hierarchical tree, the update intervals in online transformation, and the comparison of recognition rates with online adaptation proposed by Huo and Lee [16] in incremental mode as well as batch mode.

A. OLT Evaluated by Tree Construction Using Various Distance Measures

In this study, we need to build a hierarchical tree recording the cluster memberships of HMM units in different tree levels. When the adaptation data are presented sequentially, we can effectively estimate the transformation parameters of all HMM units for any amount of adaptation data. Herein, three distance

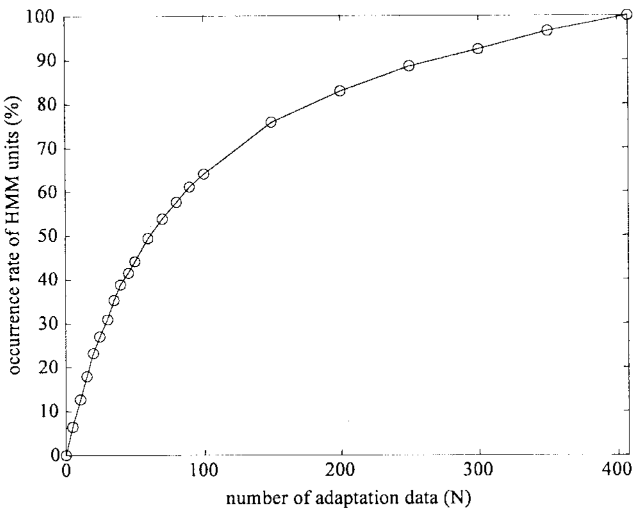

Fig. 3. Occurrence rate of HMM units versus number of adaptation data.

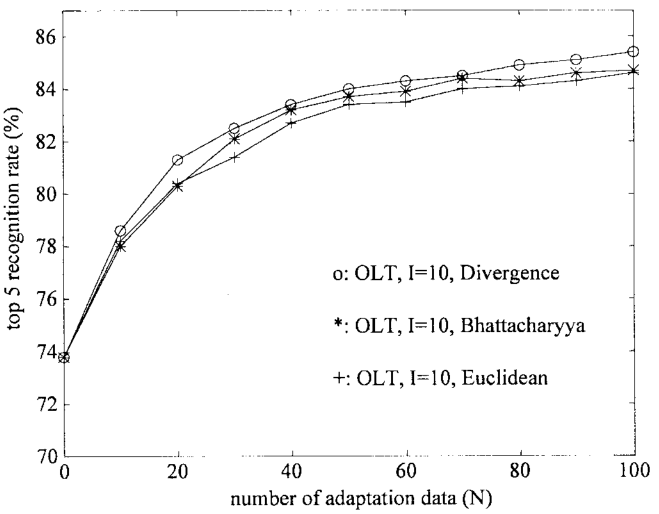

Fig. 4. Comparison of top five recognition rates of online transformation with hierarchical clustering by using various distance measures. Update interval is ten utterances.

measures are investigated in building the hierarchical tree. To evaluate the effects of various distance measures, we compare their recognition rates by using the online transformed speech models. In this set of experiments, we carry out the top five recognition rates for different numbers of adaptation data. The number of adaptation data is evaluated once every ten utterances (i.e., update interval (I) is ten utterances). After each evaluation, we discard the current data segment and use next data segment to perform online transformation (OLT). The result of $N = 0$ corresponds to a baseline SI result. As shown in Fig. 4, no matter what distance measure is applied, the recognition performance is improved continuously and consistently for increasing number of adaptation data. This proves the convergence property of proposed method. On the other hand, we can see that the performances obtained by three distance measures are comparable though the use of divergence measure gives a little improvement. In case of $N = 100$, the top five recognition rate has been raised to 85.4%. In the experiments reported later, we only carry out the results by using divergence measure.

B. OLT Evaluated by Hierarchical Tree with Various Depths

According to online hierarchical transformation, the speaker adaptation is performed by the aid of HMM's hierarchical structure. The results given in Fig. 4 were obtained by using a hierarchical tree with eight layers. In this set of experiments, we would investigate the effects of hierarchical tree with various depths. In case of six tree layers, the transformation parameters are calculated for six layers at most. As described in Section III-C, when fewer hierarchical layers are applied, the locality of model transformation becomes vaguer. Fig. 5 shows the top five recognition rates of on-line transformation with various tree depths. Herein, the update interval is also ten utterances $(I = 10)$. From the figure, we find that online transformation using eight layers achieves the best results. In case of $N = 100$, the top five recognition rates are 81.4%, 84.1%, and 85.4% for tree depths being three, six, and eight, respectively. These results further confirm the effectiveness of a deep hierarchical tree. Thus, we fix the depth of hierarchical tree to be eight in the following evaluations.

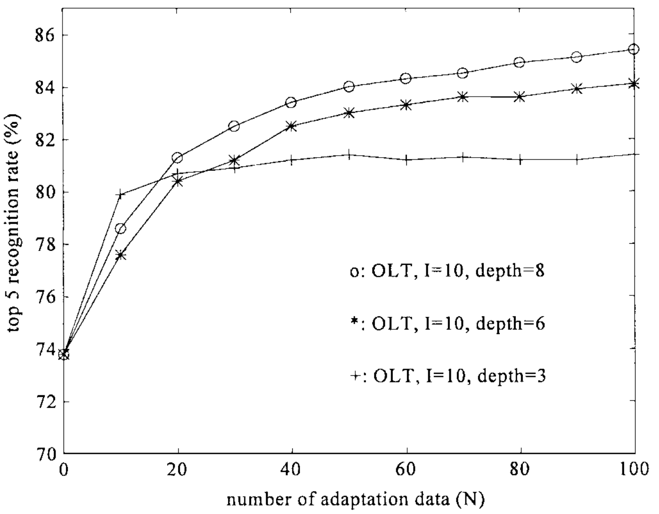

Fig. 5. Comparison of top five recognition rates of online transformation with various depths in hierarchical tree. Update interval is ten utterances.

C. OLT Evaluated by Various Update Intervals

Basically, the main advantage of online model adaptation lies in the convenience of data collection for model adaptation. Using limited interval of speech data and hyperparameters of transformation functions, the HMM parameters can be incrementally and effectively transformed to a new speaker. In this set of experiments, we focus on the evaluation of various update intervals in OLT. Herein, the total number of adaptation data is fixed at $N = 100$. The update intervals of $I = 5$, 10, 20, 25, 50 and 100 are considered in this comparative study. Note that the case of $I = 100$ (i.e., update interval equals to number of adaptation data) corresponds to perform the *batch adaptation*. As demonstrated in Fig. 6, the

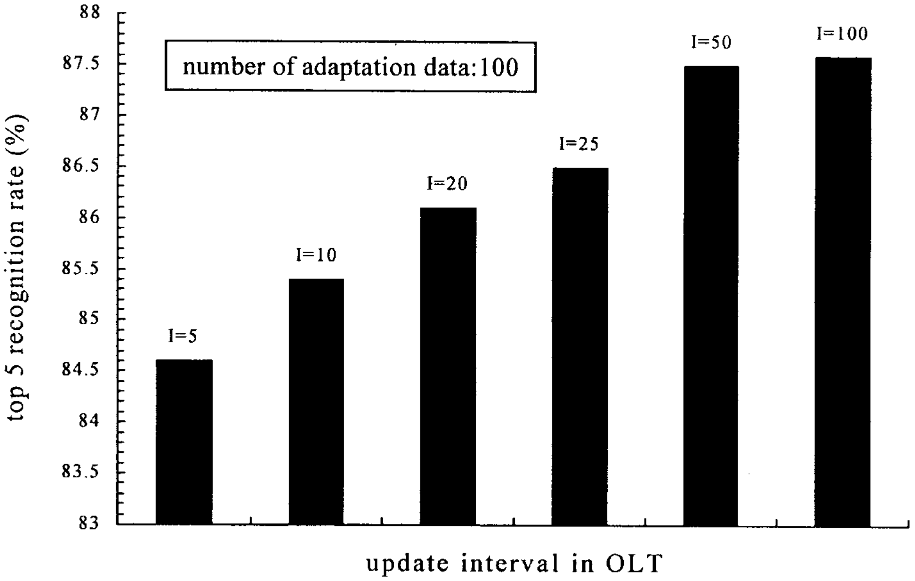

Fig. 6. Comparison of top five recognition rates of online transformation with various update intervals. Total number of adaptation data is 100.

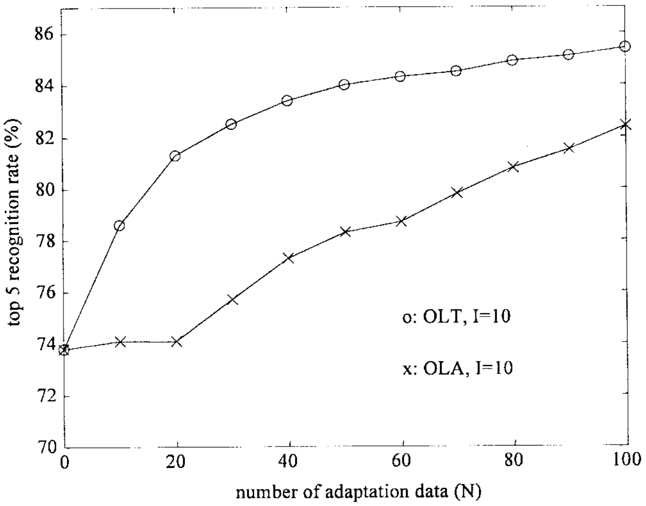

Fig. 7. Comparison of top five recognition rates of online transformation and Huo's online adaptation. Update interval is ten utterances.

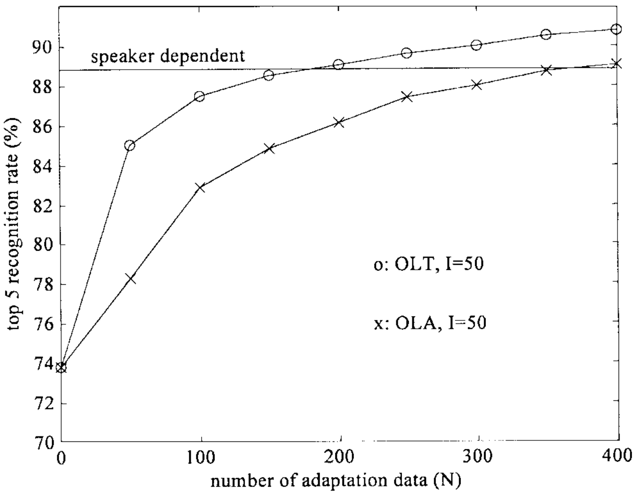

Fig. 8. Comparison of top five recognition rates of online transformation, Huo's online adaptation and SD result. Update interval is 50 utterances.

top five recognition rates are increased from 84.6% of $I = 5$ to 87.6% of $I = 100$. This is because that long interval of speech data contains sufficient knowledge of training tokens and phonetic units. The goodness of estimated transformation parameters could be guaranteed. However, a long interval period of data collection is usually undesirable in the real world due to high computational cost and expensive memory storage. Therefore, it is a tradeoff between update interval and recognition performance in OLT.

D. Comparison of OLT and OLA in Incremental Mode and in Batch Mode

As mentioned in preceding descriptions, the superiority of proposed OLT over Huo's online adaptation (OLA) [16] is the capability of overall transformation of HMM units even though most of sounds are unheard in adaptation data. The OLT is capable of transforming all HMM units effectively by employing the HMM's hierarchy. Conversely, the OLA in [16] only adjusts the HMM units appearing in adaptation data. The other HMM units are kept unchanged (i.e., using SI models). In this set of evaluation, we compare the performances of OLT and OLA in terms of incremental mode as well as batch mode. First of all, two sets of experiments in incremental mode are reported and illustrated in Figs. 7 and 8. In Fig. 7, the maximum number of adaptation data and the update interval are selected to be $N = 100$ and $I = 10$, respectively. In Fig. 8, we investigate the effects of larger amount of adaptation data with longer update interval, i.e., $N = 400$ and $I = 50$. Also, the speaker dependent (SD) performance is included in Fig. 8 for comparison. Herein, the SD result is obtained by using SD models trained

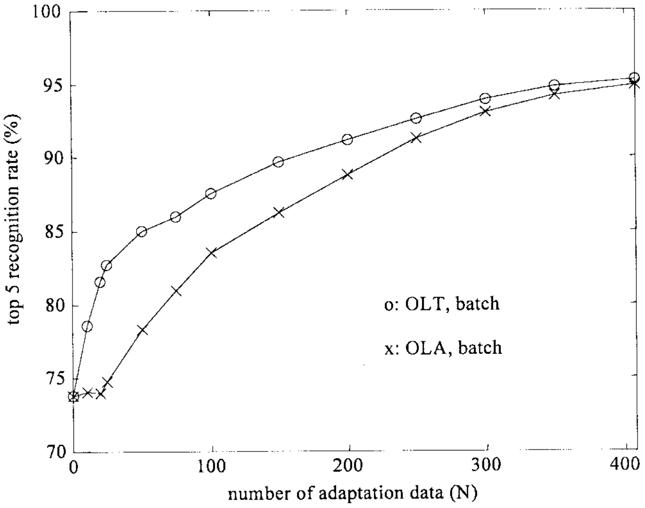

Fig. 9. Comparison of top five recognition rates of batch versions of online transformation and Huo's on-line adaptation.

from 400 adaptation utterances of the enrolled speaker. In case of $I = 10$, the occurrence rate of HMM units in each incremental segment is only about 12.6% (as shown in Fig. 3). It means that 87.4% of HMM units can not be adapted by using OLA although more HMM units may be observed and adapted later for larger amount of data. However, OLT has the ability to extract the transformation parameters of all HMM units. When the occurrence rate of collected data is increased continuously, the OLT can dynamically select the estimated transformation factors or the updated hyperparameters for model transformation. From both figures, we observe the consistent improvement of OLA as data amount N is increased. Besides, the recognition performance of OLT is significantly better than that of OLA for both small N and large N. This proves the effectiveness and efficiency of proposed OLT. The improvement is especially obvious for small N. For examples, in case of $I = 10$, the top five recognition rate of OLT at $N = 20$ is 81.3% which is excellent compared with 74.1% of OLA. Even in case of $I = 50$ (occurrence rate in each update interval is about 44.2%), the OLT at $N = 400$ (i.e., almost all HMM units are seen) has a recognition rate of 90.8% which is better than 89% of OLA. The reason is that OLA can not perform the adaptation of all HMM parameters within an update interval of 50 utterances. Furthermore, we find that OLA approaches to SD result (88.8%) at $N = 400$. But, the proposed OLT is substantially better than SD result.

On the other hand, the batch adaptation of OLT and OLA is our another concerning issue. As addressed in Section IV-C, the batch adaptation is a special case of online adaptation when the number of adaptation data equals to the update interval. In theory, the batch mode of Huo's OLA is analogous to the MAP framework of [14]. In our previous works [4], [5], the MAP batch adaptation based on the transformation function defined in (7) was also addressed. It can be viewed as the batch version of proposed OLT. Fig. 9 compares the top five recognition rates obtained by batch versions of OLT and OLA. Again, we find that OLT outperforms OLA in batch mode especially for small N. As an example, the recognition rate by using batch OLA at $N = 25$ is only 74.8%. Using batch OLT, the recognition rate can be greatly improved to 82.7%. When sufficient adaptation data (for example $N = 408$) is employed, batch OLT and batch OLA achieve comparable results. It is not surprising because that the batch OLA has nice asymptotic characteristic for large N [14]. Also, the batch OLT can effectively search the transformation parameters of individual HMM units for large N due to the incorporation of HMM's structure. From these encouraging results, we conclude that the proposed online hierarchical transformation is an excellent approach to incremental adaptation in large scale's HMM-based speech recognition. The adaptation performance is good for any practical amount of adaptation data.

V. Conclusion

We have extended the framework of recursive QB estimate to learn the parameters for online transformation-based adaptation. The main purpose of proposed method is to incrementally trace the acoustical variabilities and adapt the model parameters to fit the newest variabilities without the need of storing previous consecutive data. In this study, we emphasize our contribution on the development of online transformation of overall HMM parameters in large-vocabulary speech recognition system. Our method is to apply the prior knowledge of HMM structure and estimate the parameters of structural nodes for online transformation. This method is really adaptive in nature for speech recognition. In a series set of speaker adaptation experiments, we compared the proposed method with Huo's online adaptation [16] and evaluated the effects of various hierarchical structures and update intervals in online transformation. The conclusions are summarized as follows.

1) The performance of proposed online transformation is improved consistently as number of adaptation data is increased.
2) The use of hierarchical tree is important for online transformation. Building tree structure based on divergence measure is a good choice. When online transformation is performed, the use of deeper hierarchical tree makes the transformation more locally and therefore obtains better recognition results.
3) In online transformation, longer update interval provides larger training tokens and phonetic units to achieve higher recognition rates. The memory and computational requirements are increased as well.
4) Due to the capability of transforming all HMM units by using insufficient data, the proposed online transformation is superior to Huo's online adaptation [16] for various update intervals and data amounts. In terms of batch mode, online transformation outperforms Huo's online adaptation for limited adaptation data and approaches to Huo's online adaptation for abundant adaptation data.

APPENDIX
DERIVATION OF THE EXPECTATION AND MAXIMIZATION STEPS

Under the specification of prior density in (8), the auxiliary function in expectation step is expanded by

$$\begin{aligned}R\big(\hat{\eta}_c^{(n)}\,\big|\,\eta_c^{(n)}\big) &= R\big(\hat{\mu}_c^{(n)},\hat{\theta}_c^{(n)}\,\big|\,\mu_c^{(n)},\theta_c^{(n)}\big)\\ &= \sum_t \sum_{i,k\in\Omega_e} \xi_t(i,k)\bigg[\frac{1}{2}\log\big|\hat{\theta}_c^{(n)}r_{ik}\big|\\ &\quad - \frac{1}{2}\big(\mathbf{x}_t^{(n)}-\mu_{ik}-\hat{\mu}_c^{(n)}\big)^T\\ &\quad \times\,\hat{\theta}_c^{(n)}r_{ik}\big(\mathbf{x}_t^{(n)}-\mu_{ik}-\hat{\mu}_c^{(n)}\big)\bigg]\\ &\quad + \frac{1}{2}\big(\alpha_c^{(n-1)}-d\big)\log\big|\hat{\theta}_c^{(n)}\big|\\ &\quad - \frac{1}{2}\big(\hat{\mu}_c^{(n)}-m_c^{(n-1)}\big)^T\hat{\theta}_c^{(n)}\tau_c^{(n-1)}\\ &\quad \times\,\big(\hat{\mu}_c^{(n)}-m_c^{(n-1)}\big)-\frac{1}{2}\operatorname{tr}\big(\mu_c^{(n-1)}\hat{\theta}_c^{(n)}\big)\end{aligned} \tag{31}$$

where $\xi_t(i,k)=\Pr(s_t^{(n)}=i,l_t^{(n)}=k \mid \mathbf{X}_n,\mu_c^{(n)},\theta_c^{(n)})$. We may further define the sufficient statistics of c_{ik}, sample mean $\bar{\mathbf{b}}_c$ and sample covariance S_{ik} as given in (13)–(15). Then, by using the relation

$$\begin{aligned}&\sum_t \sum_{i,k\in\Omega_c} \xi_t(i,k)\big(\mathbf{x}_t^{(n)}-\mu_{ik}-\hat{\mu}_c^{(n)}\big)^T\hat{\theta}_c^{(n)}r_{ik}\\ &\quad\times\,\big(\mathbf{x}_t^{(n)}-\mu_{ik}-\hat{\mu}_c^{(n)}\big)\\ &\qquad = \sum_{i,k\in\Omega_c}\sum_t \xi_t(i,k)\big(\hat{\mu}_c^{(n)}-\bar{\mathbf{b}}_c\big)^T\hat{\theta}_c^{(n)}r_{ik}\big(\hat{\mu}_c^{(n)}-\bar{\mathbf{b}}_c\big)\\ &\qquad\quad + \operatorname{tr}\left(\sum_{i,k\in\Omega_c}\sum_t \xi_t(i,k)\big(\mathbf{x}_t^{(n)}-\mu_{ik}-\bar{\mathbf{b}}_c\big)\right.\\ &\qquad\quad \left.\times\,\big(\mathbf{x}_t^{(n)}-\mu_{ik}-\bar{\mathbf{b}}_c\big)^T\hat{\theta}_c^{(n)}r_{ik}\right)\end{aligned} \tag{32}$$

and the property of the trace, (31) can be rearranged by

$$\begin{aligned}&R\big(\hat{\eta}_c^{(n)}\,\big|\,\eta_c^{(n)}\big)\\ &\quad = \frac{1}{2}\sum_{i,k\in\Omega_c} c_{ik}\log\big|\hat{\theta}_c^{(n)}\big| + \frac{1}{2}\sum_{i,k\in\Omega_c} c_{ik}\log|r_{ik}|\\ &\qquad + \frac{1}{2}\big(\alpha_c^{(n-1)}-d\big)\log\big|\hat{\theta}_c^{(n)}\big|\\ &\qquad - \frac{1}{2}\sum_{i,k\in\Omega_c} c_{ik}\big(\hat{\mu}_c^{(n)}-\bar{\mathbf{b}}_c\big)^T\hat{\theta}_c^{(n)}r_{ik}\big(\hat{\mu}_c^{(n)}-\bar{\mathbf{b}}_c\big)\\ &\qquad - \frac{1}{2}\big(\hat{\mu}_c^{(n)}-m_c^{(n-1)}\big)^T\hat{\theta}_c^{(n)}\tau_c^{(n-1)}\big(\hat{\mu}_c^{(n)}-m_c^{(n-1)}\big)\\ &\qquad - \frac{1}{2}\operatorname{tr}\left(\sum_{i,k\in\Omega_c} S_{ik}r_{ik}\hat{\theta}_c^{(n)}\right) - \frac{1}{2}\operatorname{tr}\big(u_c^{(n-1)}\hat{\theta}_c^{(n)}\big).\end{aligned} \tag{33}$$

However, because the fourth and fifth terms in above equation can be verified by [7]

$$\begin{aligned}&\big(\hat{\mu}_c^{(n)}-m_c^{(n-1)}\big)^T\hat{\theta}_c^{(n)}\tau_c^{(n-1)}\big(\hat{\mu}_c^{(n)}-m_c^{(n-1)}\big)\\ &\quad + \big(\hat{\mu}_c^{(n)}-\bar{\mathbf{b}}_c\big)^T\hat{\theta}_c^{(n)}\sum_{i,k\in\Omega_c} c_{ik}r_{ik}\big(\hat{\mu}_c^{(n)}-\bar{\mathbf{b}}_c\big)\\ &\quad = \big(\hat{\mu}_c^{(n)}-\hat{m}_c\big)^T\hat{\theta}_c^{(n)}\left(\tau_c^{(n-1)}+\sum_{i,k\in\Omega_c} c_{ik}r_{ik}\right)\\ &\qquad \times\,\big(\hat{\mu}_c^{(n)}-m_c\big)-\hat{m}_c^T\hat{\theta}_c^{(n)}\left(\tau_c^{(n-1)}+\sum_{i,k\in\Omega_c} c_{ik}r_{ik}\right)\hat{m}_c\\ &\qquad + m_c^{(n-1)T}\hat{\theta}_c^{(n)}\tau_c^{(n-1)}m_c^{(n-1)}+\bar{\mathbf{b}}_c^T\hat{\theta}_c^{(n)}\sum_{i,k\in\Omega_c} c_{ik}r_{ik}\bar{\mathbf{b}}_c\\ &\quad = \big(\hat{\mu}_c^{(n)}-\hat{m}_c\big)^T\hat{\theta}_c^{(n)}\hat{\tau}_c\big(\hat{\mu}_c^{(n)}-\hat{m}_c\big)\\ &\qquad + \operatorname{tr}\left(\hat{\tau}_c^{-1}\tau_c^{(n-1)}\sum_{i,k\in\Omega_c} c_{ik}r_{ik}\big(\bar{\mathbf{b}}_c-m_c^{(n-1)}\big)^T\hat{\theta}_c^{(n)}\right.\\ &\qquad \left.\times\,\big(\bar{\mathbf{b}}_c-m_c^{(n-1)}\big)\right)\\ &\quad = \big(\hat{\mu}_c^{(n)}-\hat{m}_c\big)^T\hat{\theta}_c^{(n)}\hat{\tau}_c\big(\hat{\mu}_c^{(n)}-\hat{m}_c\big)\\ &\qquad + \operatorname{tr}\left(\hat{\tau}_c^{-1}\tau_c^{(n-1)}\sum_{i,k\in\Omega_c} c_{ik}r_{ik}\big(\bar{\mathbf{b}}_c-m_c^{(n-1)}\big)\right.\\ &\qquad \left.\times\,\big(\bar{\mathbf{b}}_c-m_c^{(n-1)}\big)^T\hat{\theta}_c^{(n)}\right)\end{aligned} \tag{34}$$

accordingly, the exponent of the auxiliary function in (33) multiplied by a constant, $K\cdot\exp\{R(\hat{\eta}_c^{(n)} \mid \eta_c^{(n)})\}$, can be expressed in the following normal-Wishart density

$$\begin{aligned}&\exp\left\{R\big(\hat{\eta}_c^{(n)}\,\big|\,\eta_c^{(n)}\big)-\frac{1}{2}\sum_{i,k\in\Omega_c} c_{ik}\log|r_{ik}|\right\}\\ &\quad = K\cdot\exp\big\{R\big(\hat{\eta}_c^{(n)}\,\big|\,\eta_c^{(n)}\big)\big\} = g\big(\hat{\mu}_c^{(n)},\hat{\theta}_c^{(n)}\,\big|\,\hat{\varphi}_c\big)\\ &\quad \propto \big|\hat{\theta}_c^{(n)}\big|^{(\hat{\alpha}_c-d)/2}\exp\left[-\frac{1}{2}\big(\hat{\mu}_c^{(n)}-\hat{m}_c\big)^T\hat{\theta}_c^{(n)}\hat{\tau}_c\big(\hat{\mu}_c^{(n)}-\hat{m}_c\big)\right]\\ &\qquad \times\exp\left[-\frac{1}{2}\operatorname{tr}\big(\hat{\mu}_c\hat{\theta}_c^{(n)}\big)\right].\end{aligned} \tag{35}$$

Herein, the new hyperparameters $\hat{\varphi}_c=(\hat{\tau}_c,\hat{m}_c,\hat{\alpha}_c,\hat{u}_c)$ are defined in (9)–(12). In the maximization step, we take the gradient of $g(\hat{\mu}_c^{(n)},\hat{\theta}_c^{(n)} \mid \hat{\varphi}_c)$ with respect to the transformation parameters $\hat{\eta}_c^{(n)}=(\hat{\mu}_c^{(n)},\hat{\theta}_c^{(n)})$. The new QB estimates of $\hat{\mu}_c^{(n)}$ and $\hat{\theta}_c^{(n)}$ are finally derived and shown in (16) and (17).

ACKNOWLEDGMENT

The author acknowledges Dr. C.-H. Lee of Bell Laboratories and Prof. Q. Huo of the University of Hong Kong for their helpful comments, and the anonymous reviewers for their critical comments.

REFERENCES

[1] V. Abrash, A. Sankar, H. Franco, and M. Cohen, “Acoustic adaptation using nonlinear transformations of HMM parameters,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1996, pp. 729–732.

[2] P. Baldi and Y. Chauvin, “Smooth on-line learning algorithms for hidden Markov models,” *Neural Comput.*, vol. 6, pp. 307–318, 1994.

[3] J.-T. Chien and H.-C. Wang, “Telephone speech recognition based on Bayesian adaptation of hidden Markov models,” *Speech Commun.*, vol. 22, pp. 369–384, Sept. 1997.

[4] J.-T. Chien, C.-H. Lee, and H.-C. Wang, "Improved Bayesian learning of hidden Markov models for speaker adaptation," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Munich, Germany, Apr. 1997, vol. 2, pp. 1027–1030.

[5] ——, "A hybrid algorithm for speaker adaptation using MAP transformation and adaptation," *IEEE Signal Processing Lett.*, vol. 4, pp. 167–169, June 1997.

[6] J.-T. Chien, H.-C. Wang, and C.-H. Lee, "Bayesian affine transformation of HMM parameters for instantaneous and supervised adaptation in telephone speech recognition," in *Proc. 5th Eur. Conf. Speech Communication and Technology*, Rhodes, Greece, Sept. 1997, vol. 5, pp. 2575–2578.

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

[8] 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.

[9] V. Digalakis, "On-line adaptation of hidden Markov models using incremental estimation algorithms," in *Proc. 5th Eur. Conf. Speech Communication and Technology*, Sept. 1997, vol. 4, pp. 1859–1862.

[10] 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.

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

[12] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic, 1990.

[13] M. J. F. Gales, "Transformation smoothing for speaker and environmental adaptation," in *Proc. 5th Eur. Conf. Speech Communication and Technology*, Sept. 1997, vol. 4, pp. 2067–2070.

[14] 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, 1994.

[15] 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. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, May 1995, pp. 457–460.

[16] 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, Mar. 1997.

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

[18] 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. 334–344, 1995.

[19] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Comput.*, vol. 6, pp. 181–214, 1994.

[20] 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, 1986.

[21] T. Kosaka, S. Matsunaga, and S. Sagayama, "Speaker-independent speech recognition based on tree-structured speaker clustering," *Comput. Speech Lang.*, vol. 10, pp. 55–74, 1996.

[22] V. Krishnamurthy and J. B. Moor, "On-line estimation of hidden Markov model parameters based on the Kullback-Leibler information measure," *IEEE Trans. Signal Processing*, vol. 41, pp. 2557–2573, 1993.

[23] 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.

[24] C.-H. Lee *et al.*, "Improved acoustic modeling for large vocabulary continuous speech recognition," *Comput. Speech Lang.*, vol. 6, pp. 103–127, 1992.

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

[26] 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.

[27] T. Matsuoka and C.-H. Lee, "A study of on-line Bayesian adaptation for HMM-based speech recognition," in *Proc. 3rd Eur. Conf. Speech Communication and Technology*, 1993, pp. 815–818.

[28] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1996, pp. 733–736.

[29] M. Padmanabhan, L. R. Bahl, D. Nahamoo, and M. A. Picheny, "Speaker clustering and transformation for speaker adaptation in speech recognition systems," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 71–77, 1998.

[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] H. Robbins, "The empirical Bayes approach to statistical decision problems," *Ann. Math. Stat.*, vol. 35, pp. 1–20, 1964.

[32] 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.

[33] K. Shinoda and T. Watanabe, "Speaker adaptation with autonomous model complexity control by MDL principle," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1996, pp. 717–720.

[34] K. Shinoda and C.-H. Lee, "Structural MAP speaker adaptation using hierarchical priors," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997.

[35] K. Shinoda and C.-H. Lee, "Unsupervised adaptation using structural Bayes approach," in *Proc. IEEE Int. Conf. Acoustic, Speech and Signal Processing*, May 1998, vol. 2, pp. 793–796.

[36] A. F. M. Smith and U. E. Makov, "A quasi-Bayes sequential procedure for mixtures," *J. R. Stat. Soc. B*, vol. 40, pp. 106–112, 1978.

[37] J. Spragins, "A note on the iterative application of Bayes' rule," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 544–549, 1965.

[38] J.-I. Takahashi and S. Sagayama, "Vector-field-smoothed Bayesian learning for incremental speaker adaptation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, May 1995, vol. 1, pp. 696–699.

[39] J.-T. Tou and R. C. Gonzalez, *Pattern Recognition Principles*. Reading, MA: Addison-Wesley, 1974.

[40] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm," *IEEE Trans. Inform. Theory*, vol. 13, pp. 260–269, 1967.

[41] 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, pp. 1652–1654, 1990.

[42] G. Zavaliagkos, R. Schwartz, and J. Makhoul, "Batch, incremental and instantaneous adaptation techniques for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 1995, pp. 676–679.

[43] Y. Zhao, "Self-learning speaker and channel adaptation based on spectral variation source decomposition," *Speech Commun.*, vol. 18, pp. 65–77, 1996.

Jen-Tzung Chien (S'97–A'98–M'99) was born in Taipei, Taiwan, R.O.C., on August 31, 1967. He received the B.S. degree in electrical engineering from the Feng-Chia University, Taichung, Taiwan, in 1989, the M.S. degree in information engineering from the Tamkang University, Taipei, in 1991, and the Ph.D. degree in electrical engineering from the National Tsing Hua University, Hsinchu, Taiwan, in 1997. His Ph.D. dissertation involved research on the speech recognition in telephone environments.

He was an Instructor in the Department of Electrical Engineering, National Tsing Hua University, in 1995. Since August 1997, he has been with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, where he is currently an Assistant Professor. During the summer of 1998, he served as a Visiting Scholar at the Speech Technology Laboratory, Panasonic Technologies Inc., Santa Barbara, CA, working on unsupervised speaker adaptation. His current research interests include statistical and adaptive signal processing, speech recognition, speaker adaptation, spoken dialogue processing, pattern recognition, neural networks, and multimodel human-computer interface.

Dr. Chien is member of the European Speech Communication Association and the Association for Computational Linguistics and Chinese Language Processing.