The HKU Scholars Hub The University of Hong Kong 香港大學學術庫

Title	Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition
Author(s)	Huo, Q; Chan, C; Lee, CH
Citation	IEEE Transactions on Speech and Audio Processing, 1995, v. 3 n. 5, p. 334-345
Issued Date	1995
URL	http://hdl.handle.net/10722/43667
Rights	Creative Commons: Attribution 3.0 Hong Kong License

Bayesian Adaptive Learning of the Parameters of Hidden Markov Model for Speech Recognition

Qiang Huo, *Member, IEEE,* Chorkin Chan, *Member, IEEE,* and Chin-Hui Lee, *Senior Member, IEEE*

Abstract—**In this paper, a theoretical framework for Bayesian adaptive training of the parameters of discrete hidden Markov model (DHMM) and of semi-continuous HMM (SCHMM) with Gaussian mixture state observation densities is presented. In addition to formulating the forward-backward MAP (maximum *a posteriori*) and the segmental MAP algorithms for estimating the above HMM parameters, a computationally efficient segmental quasi-Bayes algorithm for estimating the state-specific mixture coefficients in SCHMM is developed. For estimating the parameters of the prior densities, a new empirical Bayes method based on the moment estimates is also proposed. The MAP algorithms and the prior parameter specification are directly applicable to training speaker adaptive HMM's. Practical issues related to the use of the proposed techniques for HMM-based speaker adaptation are studied. The proposed MAP algorithms are shown to be effective especially in the cases in which the training or adaptation data are limited.**

I. Introduction

THE use of hidden Markov models (HMM's) for speech recognition has become increasingly popular in the past decade (e.g. [37]). The widespread success of the HMM framework can mainly be attributed to the existence of efficient training procedures for HMM's and the ability of the HMM to capture both the temporal and spectral variability in the speech signal. The conventional maximum likelihood (ML) based algorithms assume the HMM parameters to be fixed but unknown and the parameter estimators are derived entirely from the training observation sequences (sample information) using the Baum–Welch [2]–[4], [21], [22], [31] and the segmental maximum likelihood (or segmental k-means [39], [23]) training algorithms. There are cases in which the prior information about the HMM parameters is available. Such information may, for example, come from subject matter considerations and/or from previous experiences. The investigator may wish to use such prior information, in addition to the sample observations, in making inference about the HMM parameters. As is well known, the Bayesian inference approach provides a convenient method for combining sample observations and prior information. By assuming the HMM parameters to be random, some of the prior information about the HMM parameters can sometimes be expressed in the form of *a priori* distributions. *A posteriori* distributions can now be constructed and inference can then be made based on the posterior distributions. Consequently, the flexibility in incorporating varying amount of prior information makes the Bayesian inference procedure ideal in handling the sparse training data problem that exists in most statistical pattern recognition applications.

Recently, Bayesian adaptive learning of HMM parameters has been proposed and adopted in a number of speech recognition applications. By assuming that the set of vectors assigned to each prototype is modeled by a diagonal multivariate Gaussian density, of which the prototype is the mean, Ferretti and Scarci [9] used Bayesian estimation of mean vectors to build speaker-specific codebooks in a discrete HMM (DHMM) framework. Originated in Brown *et al.*'s preliminary effort with Bayesian estimation for speaker adaptation of continuous density HMM (CDHMM) parameters in a connected digit recognizer [6], a theoretical framework of Bayesian learning was first proposed by Lee *et al.* [27] for estimating the mean and covariance matrix parameters of a CDHMM with a multivariate Gaussian state observation density. It was then extended to handle all the parameters of a CDHMM with mixture Gaussian state observation densities [11]–[14], [26]. Two algorithms for performing Bayesian adaptive learning, namely the forward-backward MAP (maximum *a posteriori*) algorithm [13], [14], [26], and the segmental MAP algorithm [11]–[14], [26], [27] have been developed and shown to be effective for many speech recognition applications [12].

By using the same Bayesian learning framework as in [11]–[14], [26], [27], we have extended [17]–[20] the formulation to estimate parameters of DHMM's and semi-continuous HMM's (SCHMM's [15], also called tied-mixture HMM's [5]). In addition to the two above-mentioned MAP estimation algorithms, a computationally efficient segmental quasi-Bayes estimation algorithm for the mixture coefficients in SCHMM is developed [17], [20]. A new empirical Bayes method for estimating the prior density parameters based on the moment estimates is also proposed [17]–[19]. We also study practical issues related to the use of the proposed algorithms in estimating HMM parameters for speaker adaptation (SA) application. This paper investigates the problem of Bayesian adaptive learning for DHMM and SCHMM. We gather together and summarize in this paper our previous results scattered in [17]–[20] and make it more accessible to the general readership.

Manuscript received September 23, 1992; revised January 27, 1995. The associate editor coordinating the review of this paper and approving it for publication was Prof. John H. L. Hansen.

Q. Huo was with the Department of Computer Science, The University of Hong Kong, Hong Kong. He is now with ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan.

C. Chan is with the Department of Computer Science, The University of Hong Kong, Hong Kong.

C.-H. Lee is with the Speech Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974 USA.

IEEE Log Number 9413732.

1063–6676/95$04.00 © 1995 IEEE

The rest of the paper is organized as follows. After a brief introduction of the concept of the Bayesian point estimation in Section II, the formulation of MAP estimation for DHMM and SCHMM are derived, respectively, in Section III and IV. In Section V, the segmental MAP estimation of the HMM is discussed and a computationally efficient method of segmental quasi-Bayes estimation for SCHMM is presented. In Section VI, the important issue of prior density estimation is addressed and an empirical Bayes method to estimate the hyperparameters of prior density based on the moment estimate is proposed. A series of experimental results along with discussions and analyses are reported in Section VII. Finally, concluding remarks are given in Section VIII.

II. Bayesian Point Estimation

In a Bayesian approach, if θ is the random parameter vector to be estimated from a sequence of T observations $x_1, x_2, \cdots, x_T$, it is assumed that an investigator's prior knowledge about θ can be summarized in a prior probability density function (PDF) $g(\theta)$, with $\theta \in \Omega$, where Ω denotes an admissible region of the parameter space. In denoting the prior PDF $g(\theta)$, we do not explicitly show the parameters of the prior PDF (often referred to as the *hyperparameters*), which are assigned values by the investigator. For notational simplicity, we use the same symbol to denote both the random variable and the value it may assume. By the use of Bayes' theorem, the prior PDF $g(\theta)$ can be combined with the sample density function $p(x_1, x_2, \cdots, x_T|\theta)$ (which is the likelihood function if viewed as a function of θ) to yield a posterior PDF:

$$p(\theta|x_1, x_2, \cdots, x_T) = \frac{p(x_1, x_2, \cdots, x_T|\theta)g(\theta)}{\int_\Omega p(x_1, x_2, \cdots, x_T|\theta)g(\theta)d\theta} . \quad (1)$$

Such a PDF can be used to make inferences about the parameters θ. Furthermore, if an investigator has a loss function that reflects the cost of an incorrect estimation, it is generally possible to obtain an estimate, say $\hat{\theta}$, which minimizes the posterior expected loss. In this case, $\hat{\theta}$ is referred to as a Bayesian point estimator that minimizes the average risk. It is well known that the *mean* of the posterior PDF is the Bayesian point estimator given that the loss function is quadratic. On the other hand, the *mode* of the posterior PDF, usually called the modal or MAP estimator, corresponds to the special zero-one loss function case. Both the mean and the mode are reasonable candidates of the point estimate of θ [25], [7], [44]. In particular, when the prior PDF $g(\theta)$ is constant over the parameter space Ω (i.e. an *improper noninformative prior* is assumed), the MAP estimator becomes the same as the classical ML estimator.

Given the MAP formulation, three closely related issues remain to be addressed: the choice of the prior distribution family, the specification of the hyperparameters for prior densities, and the solution of the MAP estimator. In the following sections, the formulation for MAP estimation of DHMM and SCHMM are derived and the above three important issues are discussed. Whenever possible we use the same notations as in [14] for all MAP formulations.

III. MAP Estimation for Discrete HMM

In this section, we discuss the MAP estimate for discrete HMM. Consider an N-state DHMM with parameter vector $\lambda = (\pi, A, B)$, where $\pi^t = [\pi_1, \pi_2, \cdots, \pi_N]$ is the initial state probability vector, $A = [a_{ij}], i, j = 1, 2, \cdots, N$, is the transition probability matrix, and $B = [b_{jk}], j = 1, \cdots, N, k = 1, \cdots, K$, with b_{jk} being the probability of observing symbol v_k in state j. The observation symbol set is denoted as $V = \{v_1, v_2, \cdots, v_K\}$.

For simplicity, prior independence of π, A and B is assumed. The prior density for λ is then

$$g(\lambda) = g(\pi) \cdot g(A) \cdot g(B) . \quad (2)$$

If the rows of π, A and B are assumed independently distributed *a priori*, and their densities assume the form of Dirichlet distributions (sometimes called multivariate beta PDF), then $g(\lambda)$ becomes a special case of the matrix beta PDF [34]:

$$g(\lambda) = K_c \cdot \prod_{i=1}^{N} \{\pi_i^{\eta_i - 1} \cdot (\prod_{j=1}^{N} a_{ij}^{\eta_{ij} - 1}) \cdot (\prod_{k=1}^{K} b_{ik}^{\nu_{ik} - 1})\} \quad (3)$$

where K_c is a normalizing factor. $\{\eta_i\}, \{\eta_{ij}\}, \{\nu_{ik}\}$ are sets of positive parameters for the prior PDF's of π, A, and B assigned by the investigator to represent his or her prior knowledge of the parameters. Although the use of a Dirichlet prior distribution has drawn some criticism [1], it does lead to a tractable analysis. Also note that the "extended natural conjugate" prior distribution that admits nonzero correlation between the rows of A, B, and π will result in complicated formulas for the moments, etc. [34].

For an observation sequence $\mathbf{x} = (x_1, x_2, \cdots, x_T)$, let $\mathbf{s} = (s_1, s_2, \cdots, s_T)$ be the unobserved state sequence, the probability of observing the state sequence $\mathbf{s}$ is simply

$$P(\mathbf{s}|\pi, A) = \pi_{s_1} \prod_{t=2}^{T} a_{s_{t-1}s_t} . \quad (4)$$

The joint probability for observing the sequence $\mathbf{x}$ and $\mathbf{s}$ can be evaluated as

$$P(\mathbf{x}, \mathbf{s}|\lambda) = \pi_{s_1} b_{s_1}(x_1) \prod_{t=2}^{T} a_{s_{t-1}s_t} b_{s_t}(x_t) . \quad (5)$$

The probability for observing the sequence $\mathbf{x}$ is then measured by

$$P(\mathbf{x}|\lambda) = \sum_{\mathbf{s}} P(\mathbf{x}, \mathbf{s}|\lambda) \quad (6)$$

where the summation is taken over all possible state sequences. Given the observation sequence $\mathbf{x}$ and the prior density $g(\lambda)$, the MAP estimate of λ can be obtained by

$$\lambda_{MAP} = \underset{\lambda}{\operatorname{argmax}} P(\mathbf{x}|\lambda)g(\lambda) . \quad (7)$$

By viewing it as a missing data problem, as noted by Dempster *et al.* [8], the expectation-maximization (EM) algorithm can be modified to produce the MAP estimate. The EM reestimation

formulas for the three parameter sets π, A, and B are as follows (see the Appendix for a derivation of these formulas):

$$\hat{\pi}_i = \frac{e_i + \eta_i - 1}{\sum_{i=1}^{N}(e_i + \eta_i - 1)} \qquad i = 1, 2, \cdots, N \tag{8}$$

$$\hat{a}_{ij} = \frac{c_{ij} + \eta_{ij} - 1}{\sum_{j=1}^{N}(c_{ij} + \eta_{ij} - 1)} \qquad i, j = 1, 2, \cdots, N \tag{9}$$

$$\hat{b}_{jk} = \frac{d_{jk} + \nu_{jk} - 1}{\sum_{k=1}^{K}(d_{jk} + \nu_{jk} - 1)} \qquad j = 1, 2, \cdots, N; \quad k = 1, 2, \cdots, K \tag{10}$$

where

$$e_i = Pr(s_1 = i|\mathbf{x}, \lambda) \tag{11}$$

$$c_{ij} = \sum_{t=1}^{T-1} Pr(s_t = i, s_{t+1} = j|\mathbf{x}, \lambda) \tag{12}$$

$$d_{jk} = \sum_{t:x_t \sim v_k} Pr(s_t = j, x_t \sim v_k|\mathbf{x}, \lambda) \tag{13}$$

and "$x_t \sim v_k$" denotes that the observation x_t is encoded as the symbol v_k. These terms can be efficiently computed by using the forward-backward algorithm [2]. Strictly speaking, to derive the above reestimation formulas, three conditions must be obeyed: 1) $e_i + \eta_i > 1$, 2) $c_{ij} + \eta_{ij} > 1$, and 3) $d_{jk} + \nu_{jk} > 1$. Extension to the case of multiple independent observation sequences is straightforward and the formulation can be found in [16], [17].

It can be seen that the above formulation computes each MAP estimate as a weighted sum of two terms, each depending on the corresponding prior parameters and the observed data, respectively. The weights are also recomputed iteratively and depend on the hyperparameters and the data in a nonlinear fashion. Note that when the number of training samples approaches infinite, the MAP reestimation formulas approach the Baum–Welch ones that are used to get an approximate ML estimate. Thus, an asymptotical similarity of the two estimates is demonstrated. Iterative use of these reestimation formulas will give estimates of the HMM parameters corresponding to a local maximum of the posterior density, provided the iterative sequence is not trapped at some saddle point, in which case, a small random perturbation of λ away from the saddle point will hopefully set the EM algorithm free. The reader is referred to a detailed account of the convergence properties of the EM algorithm in a general setting given by Wu [45]. The choice of initial estimate is therefore essential for finding a "good" solution and minimizing the number of EM iterations needed to attain a local maximum. One reasonable choice of the initial estimate is the mode of the prior density:

$$\pi_i^{(0)} = \frac{\eta_i - 1}{\sum_{i=1}^{N}(\eta_i - 1)} \qquad i = 1, 2, \cdots, N \tag{14}$$

$$a_{ij}^{(0)} = \frac{\eta_{ij} - 1}{\sum_{j=1}^{N}(\eta_{ij} - 1)} \qquad i, j = 1, 2, \cdots, N \tag{15}$$

$$b_{jk}^{(0)} = \frac{\nu_{jk} - 1}{\sum_{k=1}^{K}(\nu_{jk} - 1)} \qquad j = 1, 2, \cdots, N; \quad k = 1, 2, \cdots, K. \tag{16}$$

Note again that the following three conditions must be obeyed: 1) $\eta_i > 1$, 2) $\eta_{ij} > 1$, and 3) $\nu_{jk} > 1$. Another choice for the initial values is the mean of the prior density computed as:

$$\pi_i^{(0)} = \frac{\eta_i}{\sum_{i=1}^{N}\eta_i} \qquad i = 1, 2, \cdots, N \tag{17}$$

$$a_{ij}^{(0)} = \frac{\eta_{ij}}{\sum_{j=1}^{N}\eta_{ij}} \qquad i, j = 1, 2, \cdots, N \tag{18}$$

$$b_{jk}^{(0)} = \frac{\nu_{jk}}{\sum_{k=1}^{K}\nu_{jk}} \qquad j = 1, 2, \cdots, N; \quad k = 1, 2, \cdots, K. \tag{19}$$

Both are some kind of summarization of the available information about the parameters before any data are observed.

IV. MAP Estimation for Semi-continuous HMM

Semi-continuous [15] or tied mixture [5] HMM have been used extensively in modeling speech for recognition. In this section, we discuss the MAP estimate for SCHMM. Consider an N-state SCHMM with parameter vector $\lambda = (\pi, A, \theta)$, where π is the initial state distribution, A is the state transition matrix, and θ is the parameter vector composed of mixture parameters $\theta_i = \{\omega_{ik}, m_k, r_k\}_{k=1,2,\cdots,K}$ for each state i with the state observation PDF being a mixture of a *common* set of Gaussian PDF's shared by all the HMM states. For state i, its observation PDF has the form of

$$p_i(x_t|\theta_i) = \sum_{k=1}^{K} \omega_{ik} f_k(x_t) = \sum_{k=1}^{K} \omega_{ik} \mathcal{N}(x_t|m_k, r_k) \tag{20}$$

where $\mathcal{N}(x|m_k, r_k)$ is the kth normal mixand denoted by

$$\mathcal{N}(x|m_k, r_k) \propto |r_k|^{1/2} \exp[-\frac{1}{2}(x - m_k)^t r_k (x - m_k)] \tag{21}$$

with m_k being the D-dimensional mean vector and r_k being the $D \times D$ precision (inverse covariance) matrix. Here " $\propto$ " denotes proportionality and $|r|$ denotes the determinant of the matrix r. Each state observation density differs from another by its corresponding mixture coefficients, ω_{ik}, which satisfy the constraint $\sum_{k=1}^{K}\omega_{ik} = 1$. By combining the MAP formulations for CDHMM and for DHMM, MAP estimation for SCHMM can be derived. We will only highlight some important points here. A detailed derivation can be found in [16] and [17].

Consider a collection of M SCHMM's and a collection of K common Gaussian component densities. We have a parameter set $\Lambda = (\lambda_1, \lambda_2, \cdots, \lambda_M; \phi_1, \phi_2, \cdots, \phi_K)$, where $\lambda_m = (\pi_i^{(m)}, a_{ij}^{(m)}, \omega_{ik}^{(m)})$ denotes the set of parameters of the mth SCHMM used to characterize the mth speech unit, and $\phi_k = (m_k, r_k)$ denotes the mean vector and the precision matrix of the kth Gaussian component. For the general case in which both the mean and precision parameters are assumed random, the prior PDF of Λ is assumed to be the product of the

conjugate priors of the *complete data* for the individual HMM parameter sets as:

$$g(\Lambda) = \prod_{m=1}^{M} g(\lambda_m) \prod_{k=1}^{K} g(m_k, r_k) \tag{22}$$

where

$$g(\lambda_m) \propto \prod_{i=1}^{N} \cdot \left\{ [\pi_i^{(m)}]^{\eta_i^{(m)}-1} \cdot (\prod_{j=1}^{N} [a_{ij}^{(m)}]^{\eta_{ij}^{(m)}-1}) \cdot (\prod_{k=1}^{K} [\omega_{ik}^{(m)}]^{\nu_{ik}^{(m)}-1}) \right\} \tag{23}$$

takes the special form of a matrix beta PDF with sets of positive hyperparameters of $\{\eta_i^{(m)}\}, \{\eta_{ij}^{(m)}\}, \{\nu_{ik}^{(m)}\}$. If the Gaussian mixture component has a full precision matrix, then $g(m_k, r_k)$ is assumed to be a normal-Wishart density [7], [27], [14] of the form:

$$\begin{aligned} g(m_k, r_k) \propto\ & |r_k|^{(\alpha_k - D)/2} \\ & \cdot \exp[-\frac{\tau_k}{2}(m_k - \mu_k)^t r_k (m_k - \mu_k)] \\ & \cdot \exp[-\frac{1}{2}\mathrm{tr}(u_k r_k)] \end{aligned} \tag{24}$$

where $\{\tau_k, \mu_k, \alpha_k, u_k\}$ are the hyperparameters of the prior density such that $\alpha_k > D - 1$, $\tau_k > 0$, μ_k is a vector of dimension D and u_k is a $D \times D$ positive-definite matrix. Here $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. On the other hand, if the Gaussian mixture component has a diagonal precision matrix, then $g(m_k, r_k)$ is assumed to be a product of normal-gamma densities [7], [27], [14] with the form:

$$\begin{aligned} g(m_k, r_k) \propto\ & \prod_{d=1}^{D} r_{kd}^{(\alpha_{kd} - 1/2)} \\ & \cdot \exp[-\frac{1}{2}\tau_{kd} r_{kd} (m_{kd} - \mu_{kd})^2] \\ & \cdot \exp[-\beta_{kd} r_{kd}] \end{aligned} \tag{25}$$

where the hyperparameters τ_{kd}, α_{kd}, $\beta_{kd} > 0$, $d = 1, 2, \cdots, D$.

Let $\mathbf{x}^{(m,n)}$ denote the nth training observation sequence of length $T^{(m,n)}$ associated with the mth speech unit, and each unit has W_m such observation sequences. Let $\mathbf{s}^{(m,n)}$ denote the unobserved state sequence and $\mathbf{l}^{(m,n)}$ is the sequence of the unobserved mixture component labels corresponding to the observation sequence $\mathbf{x}^{(m,n)}$. Given the set of observation sequences $\{\mathbf{x}^{(m,n)}\}$ and the above prior PDF $g(\Lambda)$, the MAP estimates of Λ can be obtained by

$$\Lambda_{MAP} = \underset{\Lambda}{\mathrm{argmax}} \{ \prod_{m=1}^{M} \prod_{n=1}^{W_m} f(\mathbf{x}^{(m,n)} | \lambda_m) \} \cdot g(\Lambda) \tag{26}$$

where $f(\mathbf{x}^{(m,n)}|\lambda_m)$ is defined similarly as in (6). The maximization of the RHS of the (26) can also be solved by using the EM algorithm. The readers are referred to [16], [17] for a detailed derivation of the related reestimation formulas. The results are summarized as follows:

$$\hat{\pi}_i^{(m)} = \frac{\eta_i^{(m)} - 1 + \sum_{n=1}^{W_m} \gamma_1^{(m,n)}(i)}{\sum_{i=1}^{N} \eta_i^{(m)} - N + \sum_{n=1}^{W_m} \sum_{i=1}^{N} \gamma_1^{(m,n)}(i)} \quad i = 1, 2, \cdots, N \tag{27}$$

$$\hat{a}_{ij}^{(m)} = \frac{\eta_{ij}^{(m)} - 1 + \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \gamma_t^{(m,n)}(i,j)}{\sum_{j=1}^{N} \eta_{ij}^{(m)} - N + \sum_{n=1}^{W_m} \sum_{j=1}^{N} \sum_{t=1}^{T^{(m,n)}} \gamma_t^{(m,n)}(i,j)} \quad i, j = 1, 2, \cdots, N \tag{28}$$

$$\hat{\omega}_{ik}^{(m)} = \frac{\nu_{ik}^{(m)} - 1 + \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \zeta_t^{(m,n)}(i,k)}{\sum_{k=1}^{K} \nu_{ik}^{(m)} - K + \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \sum_{k=1}^{K} \zeta_t^{(m,n)}(i,k)} \quad i = 1, 2, \cdots, N;\ k = 1, 2, \cdots, K \tag{29}$$

when r_k is a full precision matrix, one has

$$\hat{m}_k = \frac{\tau_k \mu_k + \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \zeta_t^{(m,n)}(k) x_t^{(m,n)}}{\tau_k + \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \zeta_t^{(m,n)}(k)} \tag{30}$$

and (31), as shown at the bottom of the page, and when r_k is a diagonal precision matrix, one has

$$\hat{m}_{kd} = \frac{\tau_{kd} \mu_{kd} + \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \zeta_t^{(m,n)}(k) x_{td}^{(m,n)}}{\tau_{kd} + \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \zeta_t^{(m,n)}(k)} \tag{32}$$

and (33), again at the bottom of the page, where

$$\begin{aligned} \gamma_t^{(m,n)}(i,j) &= Pr(s_t^{(m,n)} = i, s_{t+1}^{(m,n)} \\ &= j | \mathbf{x}^{(m,n)}, \lambda_m) \quad 1 \le t \le T^{(m,n)} - 1 \end{aligned} \tag{34}$$

$$\gamma_t^{(m,n)}(i) = Pr(s_t^{(m,n)} = i | \mathbf{x}^{(m,n)}, \lambda_m) \quad 1 \le t \le T^{(m,n)} \tag{35}$$

$$\hat{r}_k^{-1} = \frac{u_k + \tau_k (\hat{m}_k - \mu_k)(\hat{m}_k - \mu_k)^t + \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \zeta_t^{(m,n)}(k) (x_t^{(m,n)} - \hat{m}_k)(x_t^{(m,n)} - \hat{m}_k)^t}{\alpha_k - D + \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \zeta_t^{(m,n)}(k)} \tag{31}$$

$$\hat{r}_{kd}^{-1} = \frac{2\beta_{kd} + \tau_{kd} (\hat{m}_{kd} - \mu_{kd})^2 + \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \zeta_t^{(m,n)}(k) (x_{td}^{(m,n)} - \hat{m}_{kd})^2}{2\alpha_{kd} - 1 + \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \zeta_t^{(m,n)}(k)} \tag{33}$$

$$\zeta_t^{(m,n)}(i,k) = Pr(s_t^{(m,n)} = i, l_t^{(m,n)} = k|\mathbf{x}^{(m,n)}, \lambda_m) \quad 1 \le t \le T^{(m,n)} \tag{36}$$

$$\zeta_t^{(m,n)}(k) = Pr(l_t^{(m,n)} = k|\mathbf{x}^{(m,n)}, \lambda_m) \quad 1 \le t \le T^{(m,n)}. \tag{37}$$

Here, $\zeta_t^{(m,n)}(i,k)$ and $\gamma_t^{(m,n)}(i)$ can be related according to the following equation:

$$\zeta_t^{(m,n)}(i,k) = \gamma_t^{(m,n)}(i) \cdot \frac{\omega_{ik}^{(m)} \mathcal{N}(x_t^{(m,n)}|m_k, r_k)}{\sum_{k=1}^{K} \omega_{ik}^{(m)} \mathcal{N}(x_t^{(m,n)}|m_k, r_k)}. \tag{38}$$

Again, these terms can be computed efficiently by using the forward-backward algorithm [2].

The initial estimate can be chosen as the mode of the prior PDF $g(\Lambda)$: $\{\pi_i^{(m)}\}$, $\{a_{ij}^{(m)}\}$, $\{\omega_{ik}^{(m)}\}$ have the same form as (14) $\sim$ (16) in the case of DHMM, $m_k = \mu_k$, $r_k = (\alpha_k - D)u_k^{-1}$ for the case of full precision matrix, and $r_{kd} = (\alpha_{kd} - \frac{1}{2})/\beta_{kd}$ for the case of diagonal precision matrix. Another choice is the mean of the prior PDF $g(\Lambda)$: $\{\pi_i^{(m)}\}$, $\{a_{ij}^{(m)}\}$, $\{\omega_{ik}^{(m)}\}$ also have the same form as (17) $\sim$ (19), $m_k = \mu_k$, $r_k = \alpha_k u_k^{-1}$ for the case of full precision matrix, and $r_{kd} = \alpha_{kd}/\beta_{kd}$ for the case of diagonal precision matrix.

For the cases of known mean vector m_k or known precision matrix r_k, the related formulation of MAP estimate can also be similarly derived [16].

V. Segmental MAP Estimation for HMM

Analogous to the segmental k-means algorithm [39], [23], a similar optimization criterion can be considered for the MAP estimate of HMM. For an observation sequence $\mathbf{x} = (x_1, x_2, \cdots, x_T)$, let $\mathbf{s} = (s_1, s_2, \cdots, s_T)$ be the associated unobserved state sequence. By maximizing the joint posterior density of the parameters λ and state sequence $\mathbf{s}$, $p(\lambda, \mathbf{s}|\mathbf{x})$, one has

$$\tilde{\lambda} = \arg\max_{\lambda} \max_{\mathbf{s}} p(\lambda, \mathbf{s}|\mathbf{x}) = \arg\max_{\lambda} \max_{\mathbf{s}} p(\mathbf{x}, \mathbf{s}|\lambda) g(\lambda) \tag{39}$$

where $g(\lambda)$ is the prior density for parameter λ and $\tilde{\lambda}$ is called the segmental MAP estimate of λ [27], [14]. It is easy to show that by starting with any estimate $\lambda^{(l)}$, alternate maximization over $\mathbf{s}$ and λ gives a sequence of estimates with nondecreasing values of $p(\lambda, \mathbf{s}|\mathbf{x})$, i.e., $p(\lambda^{(l+1)}, \mathbf{s}^{(l+1)}|\mathbf{x}) \ge p(\lambda^{(l)}, \mathbf{s}^{(l)}|\mathbf{x})$ with

$$\mathbf{s}^{(l)} = \arg\max_{\mathbf{s}} p(\mathbf{x}, \mathbf{s}|\lambda^{(l)}) \tag{40}$$

$$\lambda^{(l+1)} = \arg\max_{\lambda} p(\mathbf{x}, \mathbf{s}^{(l)}|\lambda) g(\lambda). \tag{41}$$

The most likely state sequence $\mathbf{s}^{(l)}$ is decoded by means of the Viterbi algorithm [10]. If maximizing the RHS of (41) has no closed form solution, it can be accomplished by any hill climbing procedure that replaces $\lambda^{(l)}$ by $\lambda^{(l+1)}$ subject to the following constraint:

$$p(\mathbf{x}, \mathbf{s}^{(l)}|\lambda^{(l+1)}) g(\lambda^{(l+1)}) \ge p(\mathbf{x}, \mathbf{s}^{(l)}|\lambda^{(l)}) g(\lambda^{(l)}). \tag{42}$$

The segmental MAP algorithms for CDHMM parameters have been derived in [27], [12], [14], we now give the corresponding algorithms for estimating the parameters of DHMM's and SCHMM's in the following subsections.

A. Segmental MAP Estimate for DHMM

By applying the Viterbi algorithm to all the training data, the sets of observations associated with each HMM state on the most likely state sequence are also available. Let $n_i^{(1)}$ denote the number of observations in state i at time $t = 1$, and n_{ij} be the transition count from state i to state j in the most likely state sequences. Furthermore, let f_{jk} denote the count of observing symbol v_k in state j. It is straightforward to show that the reestimation formulas in (8) $\sim$ (10) are the closed-form solution of (41) by replacing the e_i by $n_i^{(1)}$, c_{ij} by n_{ij} and d_{jk} by f_{jk}, respectively, for each of the HMM states.

B. Segmental MAP Estimate for SCHMM

The reestimation formulas for $\{\pi_i\}$ and $\{a_{ij}\}$ are the same as those in DHMM. For each set of the HMM state mixture coefficients and the common set of the Gaussian density parameters, we replace $\gamma_t^{(m,n)}(i)$ in (38) by $\delta(s_t^{(m,n)} - i)$

$$\zeta_t^{(m,n)}(i,k) = \delta(s_t^{(m,n)} - i) \cdot \frac{\omega_{ik}^{(m)} \mathcal{N}(x_t^{(m,n)}|m_k, r_k)}{\sum_{k=1}^{K} \omega_{ik}^{(m)} \mathcal{N}(x_t^{(m,n)}|m_k, r_k)} \tag{43}$$

where $\mathbf{s}^{(m,n)}$ is the most likely state sequence corresponding to observation sequence $\mathbf{x}^{(m,n)}$, and $\delta(\cdot)$ denotes the Kronecker delta function, the reestimation formulas in (29) $\sim$ (33) can be taken as the corresponding segmental MAP reestimation formulas.

Note that in the process of segmental MAP reestimation of SCHMM parameters, the maximization over $\{\omega_{ik}^{(m)}\}$, $\{m_k\}$ and $\{r_k\}$ in (41) is usually accomplished with an EM algorithm that itself is an iterative algorithm and very time consuming. A compromise is to perform several EM iterations provided the constraint in (42) is satisfied. The optimal scheme that allows the problem to be solved in the shortest time possible is completely experiment dependent. Another possibility to solve the problem efficiently is to use the approximate solution, such as the quasi-Bayes method, proposed in next subsection.

C. Segmental Quasi-Bayes Estimate of the Mixture Coefficients

Similar to the segmental MAP algorithm, by applying the Viterbi algorithm to the training data, sets of observations (e.g., x_1, x_2, $\cdots$, x_T) associated with each HMM state can be identified. Given the sequence of observations, the updating formula for $\{\omega_{ik}\}$ corresponding to the maximization in (41) can be derived by solving the following quasi-Bayes estimation problem for a general finite mixture distribution.

Conditional on $\omega_i = (\omega_{i1}, \omega_{i2}, \cdots, \omega_{iK})$ and density functions f_1, f_2, $\cdots$, f_K, each x_n is assumed independently observed with the PDF as shown in (20). Assuming that the

prior density for ω_i has the form of a Dirichlet density

$$g(\omega_i) = D(\omega_i|\nu_{i1}^{(0)}, \cdots, \nu_{iK}^{(0)}) \propto \prod_{k=1}^{K} \omega_{ik}^{\nu_{ik}^{(0)}-1} \tag{44}$$

where $\nu_{ik}^{(0)} > 0$, for $k = 1, \cdots, K$. After observing x_1, the posterior density of ω_i becomes

$$p(\omega_i|x_1) = \sum_{k=1}^{K} p_{ik}(x_1) D(\omega_i|\nu_{i1}^{(0)}+\delta_{k1}, \cdots, \nu_{iK}^{(0)}+\delta_{kK}) \tag{45}$$

where

$$p_{ik}(x_1) = \frac{f_k(x_1)\nu_{ik}^{(0)}}{\sum_{m=1}^{K} f_m(x_1)\nu_{im}^{(0)}} \tag{46}$$

and δ_{ij} is the Kronecker delta function $\delta_{ij} = \delta(i-j)$. Many well-known approximate Bayesian learning procedures to solve this problem arise from approximating the RHS of (45) by

$$p(\omega_i|x_1) \approx D(\omega_i|\nu_{i1}^{(0)} + \Delta_{11}, \cdots, \nu_{iK}^{(0)} + \Delta_{1K}) \tag{47}$$

where the Δ_{ij}'s take values according to a specified method. Proceeding in this way, the necessary computation could be kept within reasonable bounds.

In the quasi-Bayes procedure proposed by Smith and Makov [43], [32], it is suggested that Δ_{1k} be replaced by $p_{ik}(x_1)$ shown in (46), and therefore

$$p(\omega_i|x_1) \approx D(\omega_i|\nu_{i1}^{(1)}, \cdots, \nu_{iK}^{(1)}) \tag{48}$$

where $\nu_{ik}^{(1)} = \nu_{ik}^{(0)} + p_{ik}(x_1)$. Then, subsequent updating takes place entirely within the Dirichlet family of distributions, *viz.*, $p(\omega_i|x_1, x_2, \cdots, x_n)$ is Dirichlet with parameters

$$\nu_{ik}^{(n)} = \nu_{ik}^{(n-1)} + p_{ik}(x_n) \tag{49}$$

where $\nu_{ik}^{(n-1)}$ are parameters of $p(\omega_i|x_1, x_2, \cdots, x_{n-1})$, and

$$p_{ik}(x_n) = \frac{f_k(x_n)\nu_{ik}^{(n-1)}}{\sum_{m=1}^{K} f_m(x_n)\nu_{im}^{(n-1)}} . \tag{50}$$

In the sense that the approximate posterior distribution with a mean identical to that of the true distribution, the convergence properties were established in [43].

It can be verified from the properties of the Dirichlet distribution that the (quasi-) posterior mean for ω_{ik}, after observing $x_1, x_2, \cdots, x_n$ is given by

$$\hat{\omega}_{ik}^{(n)} = \frac{\nu_{ik}^{(n)}}{\sum_{m=1}^{K} \nu_{im}^{(n)}} \tag{51}$$

and the mode of the approximate posterior density is

$$\tilde{\omega}_{ik}^{(n)} = \frac{\nu_{ik}^{(n)} - 1}{\sum_{m=1}^{K} (\nu_{im}^{(n)} - 1)} . \tag{52}$$

Both (51) and (52) can serve as the updating formula for the mixture coefficients in the segmental quasi-Bayes learning for SCHMM's. Equation (49) is used as the updating formula of the hyperparameters.

Note that when compared with the segmental MAP algorithm, the segmental quasi-Bayes method achieves its computational efficiency at the loss of guaranteeing a monotonic increasing property of the objective function, due to its approximate nature in maximizing the RHS of the equation (41). However, it will be experimentally shown in the following sections that either (51) or (52) will lead to a reasonable estimate of ω_i. Also note that the results of the above quasi-Bayes method depend on the order of the presentation of the x_i's. A natural choice is to present the x_i's in the order of their appearance in the training speech data. Another potential advantage of the segmental quasi-Bayes method over the segmental MAP one is due to its *sequential* nature in updating both the hyperparameters of the prior distribution and the SCHMM parameters. This makes the so-called *incremental* (or *on-line*) *adaptation* of the mixture coefficients a natural mode of updating the parameters. However, in this paper, only the so-called *batch* (or *block*) *adaptation* scheme is considered. The on-line adaptation formulation will be discussed elsewhere.

VI. Hyperparameters Estimation of Prior Distribution

In previous sections, the prior density $g(\lambda)$ is assumed to be a member of a preassigned family of prior distributions. In a strict Bayesian approach, the hyperparameter vector φ of this family of PDF's $\{g(\cdot|\varphi)\}$ is also assumed known based on a subjective knowledge about λ. In reality, it is difficult to possess a complete knowledge of the prior distribution. An attractive compromise between the classical non-Bayesian approach that uses no prior information and the strict Bayesian one is to adopt the empirical Bayes (EB) approach [40], [41], [33]. Here, we use a somewhat broader interpretation of the term "empirical Bayes" than what was implied by Robbins's original definition [40], [41]. When replacing φ by any estimate derived from the previous observed data, the previous data and current data are linked in the form of a two-stage sampling scheme by a common prior PDF $g(\lambda)$ of the unknown parameters λ.

Prior density estimation and the choice of density parameters depend on the particular application of interest. In speaker adaptation application presented later in this paper, prior density $g(\lambda|\varphi)$ represents the information of the variability of a certain model among a set of different speakers. Taking the empirical Bayes approach, the speaker independent (SI) training data set $\mathbf{X}$ for estimating hyperparameters φ can be divided into different subsets $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_Q$ correspond to Q different speakers or speaker groups so that each token of the speaker independent (SI) training data is associated with a speaker (group) ID information $\mathcal{C}_l$. One may use this information to estimate the corresponding HMM's $\tilde{\Lambda} = (\tilde{\lambda}_1, \tilde{\lambda}_2, \cdots, \tilde{\lambda}_Q)$ with the classical Baum–Welch or segmental k-means algorithm, or directly derive the corresponding HMM parameters at the last iteration of SI training. One then pretends to view $\{\tilde{\lambda}_i\}$ as the random observations with the density $g(\lambda)$. Generally speaking, the maximum likelihood estimation based on the marginal density $f(\mathbf{X}|\varphi)$, or for simplicity, a modified likelihood approach based on the joint PDF $f(\mathbf{X}, \tilde{\Lambda}|\varphi)$, can

be used to estimate the hyperparameters φ [33], [13], [14]. However, under the current assumptions on the form of the prior PDF $g(\cdot|\varphi)$, getting the maximum likelihood estimates of φ is nontrivial. To further simplify the problem, the method of moment is adopted in the following to estimate φ.

In the case of a DHMM where $g(\lambda)$ is assumed to have the form of (3), i.e., a matrix beta PDF, with the properties of the moments for matrix beta PDF [34], one has the following hyperparameter estimates

$$\tilde{\eta}_i = E(\pi_i)\{\frac{E(\pi_i)[1-E(\pi_i)]}{Var(\pi_i)} - 1\} \tag{53}$$

$$\tilde{\eta}_{ij} = E(a_{ij})\{\frac{E(a_{ij})[1-E(a_{ij})]}{Var(a_{ij})} - 1\} \tag{54}$$

$$\tilde{\nu}_{ik} = E(b_{ik})\{\frac{E(b_{ik})[1-E(b_{ik})]}{Var(b_{ik})} - 1\}. \tag{55}$$

Replacing $E(\pi_i)$, $Var(\pi_i)$, $E(a_{ij})$, $Var(a_{ij})$, $E(b_{ik})$, $Var(b_{ik})$ by their corresponding sample moments from random observations $\{\tilde{\lambda}_1, \tilde{\lambda}_2, \cdots, \tilde{\lambda}_Q\}$, the moment estimates of η_i, η_{ij}, ν_{ik} are thus obtained. An *ad hoc* method to estimate the hyperparameters of prior density has also been employed. Let $\hat{e}_i$, $\hat{c}_{ij}$ and $\hat{d}_{jk}$ be the respective estimated counts of related events at the last iteration of an SI training. These counts are divided by the number of training tokens for each speech unit and then plus one. The hyperparameters are then set to these values.

In the case of a SCHMM, the hyperparameters $\{\eta_i^{(m)}\}$, $\{\eta_{ij}^{(m)}\}$,$\{\nu_{ik}^{(m)}\}$ can also be estimated in a way similar to that for their counterparts in DHMM with the method of moment or simply the *ad hoc* method (the latter is adopted in the following experiments). When the Gaussian mixture component has a diagonal precision matrix and the prior density $g(m_k, r_k)$ takes the form of (25), the moment estimates of hyperparameters α_{kd}, β_{kd}, μ_{kd}, τ_{kd} are obtained as follows:

$$\tilde{\alpha}_{kd} = [E(r_{kd})]^2/Var(r_{kd}) \tag{56}$$

$$\tilde{\beta}_{kd} = E(r_{kd})/Var(r_{kd}) \tag{57}$$

$$\tilde{\mu}_{kd} = E(m_{kd}) \tag{58}$$

$$\tilde{\tau}_{kd} = \tilde{\beta}_{kd}/\{Var(m_{kd})(\tilde{\alpha}_{kd}-1)\} \tag{59}$$

where sample means of r_{kd} and m_{kd} take the values of the SI trained parameters $r_{kd}^{(SI)}$ and $m_{kd}^{(SI)}$, and the related sample variances are obtained by

$$Var(r_{kd}) = \Big\{\sum\nolimits_l c_{kl}(z_{kdl} - r_{kd}^{(SI)})^2\Big\}/\Big\{\sum\nolimits_l c_{kl}\Big\} \tag{60}$$

$$Var(m_{kd}) = \Big\{\sum\nolimits_l c_{kl}(y_{kdl} - m_{kd}^{(SI)})^2\Big\}/\Big\{\sum\nolimits_l c_{kl}\Big\} \tag{61}$$

with

$$c_{kl} = \sum_{x\in\mathbf{x}_l}\sum_t \zeta_t(k) \tag{62}$$

and

$$z_{kdl} = \frac{c_{kl}}{\sum_{x\in\mathbf{x}_l}\sum_t \zeta_t(k)\cdot(x_{td} - m_{kd}^{(SI)})^2} \tag{63}$$

$$y_{kdl} = \frac{\sum_{x\in\mathbf{x}_l}\sum_t \zeta_t(k)\cdot x_{td}}{c_{kl}}. \tag{64}$$

Note that the constraint $\alpha_{kd} > 1$ must be satisfied, otherwise $Var(m_{kd})$ does not exist. If the moment estimate of α_{kd} in (56) violates this constraint, it is arbitrarily set to 2.0 in the following experiments. For the full covariance matrix case, the prior density $g(m_k, r_k)$ has the form of (24). It is more difficult to write down a suitable number of estimating equations for the moment estimates of τ_k, α_k, μ_k, and u_k. If one considers a more restrictive prior density family by further assuming

$$\tau_k = \alpha_k = \sum\nolimits_{m=1}^{M}\sum\nolimits_{i=1}^{N}\tilde{\nu}_{ik}^{(m)} \tag{65}$$

and $\tilde{\nu}_{ik}^{(m)}$ as the moment estimates of $\nu_{ik}^{(m)}$, then the moment estimates of μ_k and u_k can be obtained as

$$\tilde{\mu}_k = E(m_k) \tag{66}$$

$$\tilde{u}_k^{-1} = \alpha_k^{-1}E(r_k) \tag{67}$$

by replacing $E(m_k)$ and $E(r_k)$ with their corresponding sample estimates.

When enough training data are available, the above method of moment will lead to a reasonable estimate of hyperparameters φ. Note that the physical meaning of the prior density $g(\lambda|\varphi)$ is application dependent. For example, in the speaker adaptation problem, $g(\lambda|\varphi)$ may be used to represent the information of the variability of a certain model among different speakers. In another application, for example, to build the context-dependent models from context-independent model, the prior density $g(\lambda|\varphi)$ will represent the variability of λ caused by different contexts. Therefore, the training data can be divided into subsets according to the context information. Further applications of this kind of Bayesian learning method to speech recognition can be found in [12].

Also note that the prior knowledge represented by $g(\lambda|\varphi)$ does not include those deterministic ones. For example, in the left-to-right HMM's, some parameters are known and fixed, and $g(\lambda|\varphi)$ will not include them. The estimation of hyperparameters φ is still an open problem and further research is thus needed. This is a key problem in making the Bayesian learning method applicable to adaptive training of HMM's.

VII. Speaker Adaptation Experiments

A. Experimental Setup

To examine the viability of the proposed techniques, the Bayesian adaptive learning framework is applied to speaker adaptation application and a series of experiments are conducted. The 26 letters of the English alphabet are chosen as the vocabulary for all experiments. Two databases are used for evaluating the adaptation algorithms, *viz.*, the OGI ISOLET and the TI46 corpora. These two databases were recorded at two separate sites with a time gap of 10 years. The sampling rates and quantization precisions are 16 KHz with 16-bit quantization and 12.5 KHz with 12-bit quantization, respectively. The speech in the ISOLET corpus is recorded with a Sennheiser HMD 224 close-talking noise-cancelling

microphone and the one in TI46 is 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 down-sampled to 8 KHz so that hopefully, they will become more compatible to each other. The feature vectors used in this study consist of 12 bandpass-liftered LPC-derived cepstral coefficients with a 30 ms frame length and a 10 ms frame shift [28]. For SI training and prior density estimation, the OGI ISOLET database is used. It consists of 150 speakers, 75 females and 75 males, each speaking each of the letters twice. For speaker dependent (SD) or adaptive training and testing, the English alphabet subset of the TI46 isolated word corpus is used. It is produced by 16 speakers, eight females and eight males. Among them, four males' data are incomplete, so only 12 speakers are used in this study. Each person utters each of the letters 26 times, 10 of them are used for SD/SA training and the remaining 16 tokens are used for testing. Throughout the experiments, each of the 26 letters in the vocabulary is modeled by a single left-to-right five-state HMM with arbitrary state skipping. In recognition, the decision rule determines the recognized letter as the one that attains the highest forward-backward probability.

B. VQ-Based Speaker Clustering

Since there are not enough data from a single speaker to estimate a model for each letter, speaker clustering based on vector quantization is performed to obtain 16 speaker clusters from which 16 sets of models needed to obtain the moment estimates of the hyperparameters are derived. To use a VQ method for speaker clustering is motivated by its simplicity and its success in speaker recognition problems [42], albeit other alternatives (e.g., [38], [35], [36], [24]). The speaker clustering process begins from two natural male/female groups. The clustering algorithm is as follows:

1) View all male speakers as one group and all female speakers as another group. Generate two codebooks of size 256, one for the female and the other for the male speaker groups, respectively.
2) Perform "speaker classification" with respect to each codebook of each speaker group with the VQ method [42].
3) Reformulate the codebook for each speaker group with the speaker classification result in Step 2.
4) If the speaker classification process is stable, a predefined maximum number of iterations is reached or the variation of the total quantization error (for all speakers) is less than a predefined threshold, then go to Step 5; else go to Step 2.
5) If a predefined number of speaker groups is reached, stop; else go to Step 6.
6) Split the codebook by a simple perturbation method, go to Step 2.

The criterion used here is to minimize the "total quantization error," so the number of speakers in each group does not have to be the same. All the training utterances of each speaker are used for "speaker classification." Each codebook is generated by using the LBG algorithm [30] with an Euclidean distortion measure. Table I shows the male/female composition of the clusters created by this algorithm. Most clusters are dominated by either male or female. In the case of two clusters, one cluster is completely male, and the other completely female. This is a very positive indication of the clustering method validity.

TABLE I
SPEAKER CLUSTERING RESULTS OF 2 TO 16 CLUSTERS

Number of Clusters		Cluster Number															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
16	Male	0	7	0	12	0	12	0	10	0	9	0	10	0	9	0	6
	Female	8	0	9	0	17	0	6	1	10	0	11	0	7	0	6	0
8	Male	0	16	0	22	0	21	0	16								
	Female	18	0	20	0	24	0	12	1								
4	Male	0	36	0	39												
	Female	43	0	31	1												
2	Male	0	75														
	Female	75	0														

TABLE II
SUMMARY OF DHMM RESULTS (SI RECOGNITION RATE: 44.1%)

No. of Tokens	1	2	3	4	5	6	7	8	9	10
SD	46.3	53.5	57.1	59.6	61.4	61.7	62.4	63.4	63.2	64.6
SA1	54.1	59.1	61.7	62.6	64.1	64.3	65.2	65.9	66.3	66.8
SA2	48.1	54.6	58.2	61.1	63.2	63.6	64.1	64.6	65.2	65.6

C. MAP Estimates of DHMM Parameters

In this subsection, experimental results concerned with the adaptation of DHMM's are discussed. A 256-vector codebook is generated from the ISOLET corpus by using the LBG algorithm [30] with a Euclidean distortion measure and is used in all subsequent experiments. The SI/SD word models are trained by using the standard Baum–Welch algorithm [2] and the SA ones are obtained by using the MAP estimate presented in Section III. The average recognition rates for 12 speakers (eight females, four males) based on the SA models are summarized in Table II. "SA1" corresponds to the speaker adaptation experiments with the *ad hoc* prior parameters, "SA2" refers to the ones with prior parameters estimated by the method of moment. For comparison purposes, the word recognition rates for the SD models (row labeled as "SD") and the SI models are also reported.

Table II clearly shows that the regular MLE training procedure ("SD") is inadequate when the amount of available training data is insufficient. The fact that the SD recognition rate using only one training token is better than that of the SI system is a good indication of the serious mismatch between SI training set and SD testing set. The results here show that speaker adaptation can be used to reduce this mismatch. The performance for "SD" improves as the number of speaker specific training tokens increases; however, it is noted that when using the same amount of training data, SA training outperforms SD training in all the cases tested. This implies that SA training utilizes training data more effectively than

TABLE III
SD PERFORMANCE AS A FUNCTION OF THE NUMBER OF TRAINING TOKENS

No. of Tokens	1	2	3	4	5	6	7	8	9	10
Rates	52.9	65.3	71.3	74.4	76.0	77.4	77.4	77.5	78.0	79.2

TABLE IV
PROGRESSIVE ADAPTATION RECOGNITION RESULTS (SI RECOGNITION RATE: 47.8%)

No. of Tokens		1	2	3	4	5	6	7	8	9	10
+SA mix. coeff.		61.7	65.2	67.5	68.1	69.4	69.5	69.9	70.1	70.7	71.0
+SA tran. prob.		62.1	65.4	67.6	68.9	69.9	69.9	70.2	70.4	70.8	71.5
SI var	+SA mean	65.8	70.5	72.7	74.1	74.5	75.2	75.0	75.7	76.1	76.6
	+SD mean	65.4	71.3	72.9	75.1	76.3	76.0	76.6	77.2	78.0	79.2
+SA var	SA mean	64.0	69.0	70.6	72.0	72.1	73.6	73.7	75.4	75.9	76.7
	SD mean	64.7	70.4	73.3	75.8	75.4	76.0	76.1	77.1	78.4	79.4

does SD training, especially when training data are insufficient. As expected, the SA performance quickly becomes equivalent to the SD performance when the number of adaptive training tokens increases. It is also noted that the performance of "SA1" is in most cases better than that of "SA2". The hyperparameters of prior distribution estimated with the *ad hoc* method seem more robust in experiments here than that estimated with the method of moment which may suffer more from the sparse training data problem.

D. MAP Estimates of SCHMM Parameters

In this subsection, the effects of the adaptation schemes presented in Section IV for estimating SCHMM parameters are examined. A common set of 256 Gaussian mixture components with diagonal covariance matrices are used in each SCHMM state. The SI/SD word models are trained by using the standard forward-backward algorithm. A pruning strategy that keeps only the top 10 mixands when computing the likelihood in (20) is used both in training and in testing. A series of recognition experiments on 12 speakers are conducted.

The first experiment is to recognize the English alphabet subset of TI46 with the SI system trained with speech tokens from OGI ISOLET. The average recognition rate is 47.8%. The second experiment is to recognize the same TI46 subset with SD systems trained with various numbers of SD tokens for each speaker and the average recognition rates for 12 speakers are tabulated in Table III.

Once again, due to the serious mismatch between SI training set and SD testing set, the SI recognition rate is inferior even to the SD recognition rate using only one training token. In the following experiments, it will be shown that the SI performance can be improved by using the speaker adapted HMM's. To examine the SA effects of the different set of model parameters, *viz.*, the mixture coefficients, the state transition probabilities, the mean vectors, and the covariance matrices of the Gaussian mixands, respectively, a series of experiments are conducted with corresponding parameters of the SI system replaced progressively and systematically by their SA counterparts. The average recognition rates are summarized in Table IV.

The first observation can be derived by comparing the first two rows of Table IV against Table III. It shows that both the SI and the SD performances are greatly improved by using the SA training for the mixture coefficients and the transition probabilities when the SD training data are limited (e.g., one token). However, by comparing rows three and four of Table IV against the first two rows, a second observation is that only using SA training of the above parameters is insufficient. Speaker dependent information on density means is very important. By additionally using SA (SD) mean vectors, the SA performances are further greatly improved and even better than the pure SD system when insufficient training data are available for SD training (in particular here, less than three tokens). As a third observation, by comparing rows four and six against rows three and five in Table IV, respectively, it is noticed that using SD means yields consistently better recognition rates irrespective to the variances and the number of tokens for SD training or adaptation. This shows that the mean vectors of the Gaussian codebook in SCHMM can represent the essential characteristics of different speakers and can be rapidly and sufficiently estimated even with limited amount of training data by regular SD training. However, the variances cannot be reliably estimated with limited training data. As a fourth observation, in the particular setup here, by comparing rows five and six against rows three and four in Table IV, it is found that using SA variances has little advantage in comparison with using SI variances.

E. Segmental Quasi-Bayes Estimates of SCHMM Parameters

As is well known, the mixture coefficients are very important parameters in modeling speech unit with SCHMM. To examine the viability and effect of the segmental quasi-Bayes algorithm presented in Section V for estimating the mixture coefficients of SCHMM only, a series of comparative experiments are conducted. For simplicity, in SA/SD training, Gaussian mixture component PDF's and the transition probabilities are fixed to the SI-trained ones. In SA training, the hyperparameters of the prior distribution of the mixture coefficients are estimated with the *ad hoc* method discussed in Section VI. The average word recognition rates for 12 speakers are summarized in Table V. The columns in Table V correspond to the numbers of training tokens used for each SD and SA cases. "SEG-ML" stands for SD segmental ML training of the mixture coefficients and "SEG-MAP" corresponds to its MAP counterpart. "SEG-QB" stands for SA segmental quasi-Bayes training of the mixture coefficients. As expected, many facts observed in the previous subsection are repeated here. Apart from those facts, another more important observation—that the recognizer performance with the segmental quasi-Bayes method is not much different from that with the segmental MAP method—shows the viability of the quasi-Bayes approximation in maximizing the RHS of (41).

Although the *batch* (or *block*) *adaptation* scheme is adopted in this study, in view of its *sequential* nature in updating both the hyperparameters of the prior distribution and the mixture coefficients themselves, the segmental quasi-Bayes

TABLE V
SUMMARY OF RECOGNITION RESULTS WITH SEGMENTAL TRAINED SCHMM'S

No. of Tokens	1	2	3	4	5	6	7	8	9	10
SEG-ML	76.3	62.5	65.9	67.0	68.4	68.3	68.6	69.4	70.3	70.7
SEG-MAP	64.5	65.1	67.2	68.0	69.1	68.9	69.1	69.7	70.5	70.8
SEG-QB	62.0	65.0	66.8	67.8	69.0	69.0	69.3	70.0	70.1	70.4

method presented in this paper will also be very suitable for performing the *incremental* (or *on-line*) *adaptation* of the mixture coefficients in SCHMM. The segmental quasi-Bayes method presented in this paper can only be theoretically justified in the case of fixed mixture components, but adjustable coefficients. More theoretical research is definitely needed to extend this framework to cases involving adjustable mixture coefficients as well as mixture component PDFs' parameters. Before that, some pragmatic procedures that combine the quasi-Bayes adaptation of mixture coefficients and different adaptation schemes of the component density should also be experimentally tested. Research along this line of thought has been conducted. We will report the related results elsewhere.

F. General Discussions

The effects of SA training, in the particular setup here, are not so significant. This is caused by the serious mismatch between the two corpora. After more detailed analysis, it is found that the SA effects are very different among different speakers. In the Bayesian learning framework, one hopes to use prior distribution of HMM parameters to represent the information of the variability of a certain model among the speakers. If a new speaker happens to be an outlier in this prior distribution, one may get little benefit from the SA training. If the SI task is severely mismatched with the SD one, the SA training may deteriorate the performance of the SD system (this is equivalent to bringing in some abnormal training samples for SD training), but it still improves the SI system tremendously. So the SA effects depend heavily on the suitability of the prior distribution to the new speaker. To cope with the mismatch problem between the prior distribution and the new adaptation data, some kind of speaker normalization (or signal space equalization) should be performed first in the acoustic (feature) space before the Bayesian framework is applied to adapt the model parameters. In the process of model adaptation, to get a better match of prior distribution and the adaptation data, multiple set of prior distributions can be used by clustering the training data for prior distribution estimation, provided enough training data are available. Results reported in previous subsections are obtained with only a single set of SI seed models. To substantiate the above argument, by using two sets of gender-dependent seed models, recognition results corresponding to SA mixture coefficients and transition matrices with SD means and SI/SA variances in SCHMM are listed in Table VI for comparison purposes. The average SI recognition rate increases to 51.3%. As for the SA system performance, in comparison with their counterparts in the case of a single set of prior distributions, better performance is achieved with two sets of gender-dependent seed models.

TABLE VI
RESULTS WITH GENDER-DEPENDENT SEED MODELS (SI RECOGNITION RATE: 51.3%)

No. of Tokens	1	2	3	4	5	6	7	8	9	10
SD mean, SI var	70.5	74.6	76.8	77.9	78.7	79.5	80.1	81.0	81.0	81.8
SD mean, SA var	69.8	74.5	76.2	77.1	78.7	77.6	78.5	79.5	79.7	80.2

VIII. SUMMARY

In this paper, a theoretical framework for Bayesian adaptive training of the parameters of DHMM and of SCHMM with Gaussian mixture state observation densities is presented. In addition to formulating the forward-backward MAP and the segmental MAP algorithms for estimating the above HMM parameters, a computationally efficient segmental quasi-Bayes algorithm for estimating the state-specific mixture coefficients in SCHMM is developed. For estimating the parameters of the prior densities, a new empirical Bayes method based on the moment estimates is also proposed. The MAP algorithms and the prior parameter specification are directly applicable to training speaker adaptive HMM's. Practical issues related to the use of the proposed techniques for HMM-based speaker adaptation are studied. The proposed MAP algorithms are shown to be effective, especially in the cases that the training or adaptation data are limited. The MAP method is also applicable to other problems in HMM training for speech recognition such as sequential training, context adaptation, and parameter smoothing.

However, some topics in Bayesian adaptive learning of the HMM parameters still deserve further research. The most immediate one is the definition of the prior distribution and the related hyperparameters estimation problem. The segmental quasi-Bayes learning method that can be used to update the hyperparameters of the prior distribution and the HMM parameters incrementally for both mixture coefficients and the mixture component parameters in SCHMM is another topic of particular importance. When the Bayesian learning framework is applied to cope with the possible mismatch problem between training and testing conditions, the choice of the appropriate prior distribution is critical to the success of the algorithm. The lessons we've learned are that in order to handle severely mismatched cases effectively, different sources of variations should be identified and then different strategies be adopted to cope with these variations. For example, in the speaker adaptation application discussed in this study, if a speaker normalization step is first taken in the entire feature vector space before the Bayesian learning framework is applied, more significant performance improvements can then be expected. Research along this line of thought is in progress.

APPENDIX
DERIVATION OF MAP ESTIMATE FOR DHMM

Let $\mathbf{y} = (\mathbf{x}, \mathbf{s})$ denote the complete data, where $\mathbf{x}$ is the observed data and $\mathbf{s}$ the missing one. Then, the complete-data log-likelihood is

$$\log P(\mathbf{x}, \mathbf{s}|\lambda) = \log \pi_{s_1} + \sum_{t=2}^{T} \log a_{s_{t-1}s_t} + \sum_{t=1}^{T} \log b_{s_t}(x_t) \,. \tag{68}$$

As noted by Dempster *et al.* [8] and similar to [14], we define a modified auxiliary function $R(\hat{\lambda}|\lambda) = Q(\hat{\lambda}|\lambda) + \log g(\hat{\lambda})$ for a given preliminary estimate λ, where $Q(\hat{\lambda}|\lambda)$ is the auxiliary function for the E-step in ML estimation (e.g., [29]):

$$Q(\hat{\lambda}|\lambda) = E[\log P(\mathbf{x}, \mathbf{s}|\hat{\lambda})|\mathbf{x}, \lambda] \tag{69}$$

$$= \sum_{\mathbf{s}} \{\frac{P(\mathbf{x}, \mathbf{s}|\lambda)}{P(\mathbf{x}|\lambda)} \log P(\mathbf{x}, \mathbf{s}|\hat{\lambda})\} \tag{70}$$

$$= \sum_{i=1}^{N} e_i \log \hat{\pi}_i + \sum_{i=1}^{N}\sum_{j=1}^{N} c_{ij} \log \hat{a}_{ij} + \sum_{j=1}^{N}\sum_{k=1}^{K} d_{jk} \log \hat{b}_{jk} \; . \tag{71}$$

By combining (3) and (71), the modified auxiliary function can be evaluated as

$$R(\hat{\lambda}|\lambda) = Q(\hat{\lambda}|\lambda) + \sum_{i=1}^{N} (\eta_i - 1) \log \hat{\pi}_i + \sum_{i=1}^{N}\sum_{j=1}^{N} (\eta_{ij} - 1) \log \hat{a}_{ij} + \sum_{j=1}^{N}\sum_{k=1}^{K} (\nu_{jk} - 1) \log \hat{b}_{jk} + \log K_c \tag{72}$$

where K_c is just a function of $\{\eta_i\}$, $\{\eta_{ij}\}$, and $\{\nu_{ik}\}$, not dependent on $\hat{\lambda}$. By choosing $\hat{\lambda}$ to maximize the RHS of (72), the EM reestimation formulas in (8) $\sim$ (10) can thus be derived.

REFERENCES

[1] J. Aitchison and S. M. Shen, "Logistic-normal distributions: Some properties and uses," *Biometrika*, vol. 67, pp. 261–272, 1980.

[2] L. E. Baum, "An inequality and associated maximization techniques in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, pp. 1–8, 1972.

[3] L. E. Baum and J. A. Egon, "An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology," *Bull. Amer. Math. Soc.*, vol. 73, pp. 360–363, 1967.

[4] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic function of Markov chains," *Annals Math. Statis.*, vol. 41, no. 1, pp. 164–171, 1970.

[5] J. R. Bellegarda and D. Nahamoo, "Tied mixture continuous parameter modeling for speech recognition," *IEEE Trans Acoust., Speech, Signal Processing*, vol. 38, no. 12, pp. 2033–2045, 1990.

[6] P. F. Brown, C.-H. Lee, and J. C. Spohrer, "Bayesian adaptation in speech recognition," in *Proc. ICASSP-83*, Boston, May 1983, pp. 761–764.

[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," *Royal Statist. Soc., J., Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.

[9] M. Ferretti and S. Scarci, "Large-vocabulary speech recognition with speaker-adapted codebook and HMM parameters," in *Proc. Eurospeech89*, Paris, Sept. 1989, pp. 154–156.

[10] G. D. Forney, "The Viterbi algorithm," *Proc. IEEE*, vol. 61, pp. 268–278, 1973.

[11] J.-L. Gauvain and C.-H. Lee, "Bayesian learning of Gaussian mixture densities for hidden Markov models," in *Proc. DARPA Speech Natural Language Workshop*, Feb. 1991, pp. 272–277.

[12] ——, "Bayesian learning for hidden Markov model with Gaussian mixture state observation densities," *Speech Commun.*, vol. 11, pp. 205–213, 1992.

[13] ——, "MAP estimation of continuous density HMM: Theory and applications," in *Proc. DARPA Speech Natural Language Workshop*, Feb. 1992, pp. 185–190.

[14] ——, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[15] X.-D. Huang and M. A. Jack, "Semicontinuous hidden Markov models for speech signals," *Comput. Speech Language*, vol. 3, pp. 239–251, 1989.

[16] Q. Huo, "A study on several statistical acoustic modeling problems in automatic speech recognition," Ph.D. thesis, Department of Electronic Engineering & Information Science, Univ. of Science and Technology of China, Mar. 1994.

[17] Q. Huo and C. Chan, "Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition," Technical Report TR-92-08, Department of Computer Science, The Univ. of Hong Kong, Sept. 1992.

[18] Q. Huo, C. Chan, and C.-H. Lee, "Bayesian learning of the parameters of discrete and tied mixture HMM's for speech recognition," in *Proc. Eurospeech-93*, Berlin, Germany, 1993, pp. III-1567–1570.

[19] ——, "Bayesian learning of the SCHMM parameters for speech recognition," in *Proc. ICASSP-94*, Adelaide, Australia, 1994, pp. I-221–224.

[20] ——, "Segmental quasi-Bayesian learning of the mixture coefficients in SCHMM for speech recognition," in *Proc. 1994 Int. Symp. Speech, Image Processing, Neural Networks*, Hong Kong, 1994, pp. 678–681.

[21] B.-H. Juang, "Maximum-likelihood estimation of mixture multivariate stochastic observations of Markov chains," *AT&T Tech. J.*, vol. 64, no. 6, pp. 1235–1249, 1985.

[22] B.-H. Juang, S. E. Levinson, and M. M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of Markov chains," *IEEE Trans. Inform. Theory*, vol. IT-32, no. 2, pp. 307–309, Mar. 1986.

[23] B.-H. Juang and L. R. Rabiner, "The segmental k-means algorithm for estimating parameters of hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 9, pp. 1639–1641, Sept. 1990.

[24] H.-K. Kuo. "Speaker clustering with hidden Markov models," Master's thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, June 1992.

[25] I. H. LaValle, *An Introduction to Probability, Decision, and Inference*. New York:, Holt, Rinehart and Winston, 1970.

[26] C.-H. Lee and J.-L. Gauvain, "Speaker adaptation based on MAP estimation of HMM parameters," in *Proc. ICASSP-93*, Apr. 1993, pp. II-588–591.

[27] 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, no. 4, pp. 806–814, Apr. 1991.

[28] C.-H. Lee, L. R. Rabiner, R. Pieraccini, and J. G. Wilpon, "Acoustic modeling for large vocabulary speech recognition," *Comput. Speech Language*, vol. 4, pp. 127–165, 1990.

[29] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *Bell Syst. Tech. J.*, vol. 62, no. 4, pp. 1035–1074, 1983.

[30] Y. Linde, A. Buzo, and R.M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, 1980.

[31] L. R. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 729–734, 1982.

[32] U. E. Makov and A. F. M. Smith, "A quasi-Bayes unsupervised learning procedure for priors," *IEEE Trans. Inform. Theory*, vol. IT-23, no. 6, pp. 761–764, 1977.

[33] J. S. Maritz and T. Lwin, *Empirical Bayes Methods*, 2nd edition. London: Chapman and Hall, 1989.

[34] J. J. Martin, *Bayesian Decision Problems and Markov Chains*. New York: Wiley, 1967.

[35] L. Mathan and L. Miclet, "Speaker hierarchical clustering for improving speaker independent HMM word recognition," in *Proc. ICASSP-90*, Albuquerque, NM, Apr. 1990, pp. 149–152.

[36] S. Nakamura and T. Akabane, "A neural speaker model for speaker clustering," in *Proc. ICASSP-91*, Toronto, 1991, pp. 853–856.

[37] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Mar. 1989.

[38] L. R. Rabiner, C.-H. Lee, B.-H. Juang, and J.-G. Wilpon, "HMM clustering for connected word recognition," in *Proc. ICASSP-89*, Glasgow, Scotland, 1989, pp. 405–408.

[39] L. R. Rabiner, J. G. Wilpon, and B.-H. Juang. "A segmental k-means training procedure for connected word recognition," *AT&T Tech. J.*, vol. 65, no. 3, pp. 21–31, 1986.
[40] H. Robbins, "An empirical Bayes approach to statistics," in *Proc. Third Berkeley Symp. Math. Statist. Prob.*, 1955, pp. I-157–164.
[41] H. Robbins, "The empirical Bayes approach to statistical decision problems," *Ann. Math. Statist.*, vol. 35, pp. 1–20, 1964.
[42] A. E. Rosenberg and F.-K. Soong, "Evaluation of a vector quantization talker recognition system in text independent and text dependent modes," *Comput. Speech Language*, vol. 22, pp. 143–157, 1987.
[43] A. F. M. Smith and U. E. Makov, "A quasi-Bayes sequential procedure for mixtures," *Royal Statist. Soc. J., Series B*, vol. 40, no. 1, pp. 106–112, 1978.
[44] Robert L. Winkler, *Introduction to Bayesian Inference and Decision*. New York: Holt, Rinehart and Winston, 1972.
[45] C. F. Jeff Wu, "On the convergence properties of the EM algorithm," *Annals Statist.*, vol. 11, no. 1, pp. 95–103, 1983.

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

From 1986 to 1990, his research work focused on the hardware development for real-time digital signal processing, image processing and computer vision, and speech and speaker recognition.

From 1991 to 1994, he was with the Department of Computer Science, The University of Hong Kong, where he was involved in research work on speech recognition. Since April 1995, he has been with ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan. His current major research interests include speech recognition, speaker recognition, and general pattern recognition theory.

Chorkin Chan (M'81) was born and brought up in Hong Kong. After receiving the B.S. degree in mathematics from the National Taiwan University, he furthered his study in Canada and received the Ph.D. degree in physics from the University of British Columbia.

Since then, his interest has been computer science. He has worked for IBM and the University of Victoria and is currently teaching at the Department of Computer Science, University of Hong Kong. His major research interest is in speech and pattern recognition.

Chin-Hui Lee (S'78–M'82–SM'91) received the B.S. degree from National Taiwan University, Taipei, in 1973, the M.S. degree from Yale University, New Haven, CT, in 1977, and the Ph.D. degree from University of Washington, Seattle, WA, 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, where he engaged in research work in speech coding, speech recognition and signal processing for the development of the DSC-2000 Voice Server. Since 1986, he has been with AT&T Bell Laboratories, Murray Hill, NJ. His current research interests include signal processing, speech modeling, speech recognition, speaker recognition, and spoken dialogue processing.

From 1991 to 1995, he was an associate editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He is now a member of the Speech Technical Committee of the IEEE Signal Processing Society. He also serves as a member on the ARPA Spoken Language Processing Coordination Committee.