Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains

Jean-Luc Gauvain[1] **and Chin-Hui Lee**

Speech Research Department
AT&T Bell Laboratories
Murray Hill, NJ 07974

In this paper a framework for maximum a posteriori (MAP) estimation of hidden Markov models (HMM) is presented. Three key issues of MAP estimation, namely the choice of prior distribution family, the specification of the parameters of prior densities and the evaluation of the MAP estimates, are addressed. Using HMMs with Gaussian mixture state observation densities as an example, it is assumed that the prior densities for the HMM parameters can be adequately represented as a product of Dirichlet and normal-Wishart densities. The classical maximum likelihood estimation algorithms, namely the forward-backward algorithm and the segmental k-means algorithm, are expanded and MAP estimation formulas are developed. Prior density estimation issues are discussed for two classes of applications: parameter smoothing and model adaptation, and some experimental results are given illustrating the practical interest of this approach. Because of its adaptive nature, Bayesian learning is shown to serve as a unified approach for a wide range of speech recognition applications.

1 Introduction

Estimation of a probabilistic function of Markov chain, also called a hidden Markov model (HMM), is usually obtained by the method of *maximum likelihood* (ML) [1, 2, 23, 15] which assumes that the size of the training data is large enough to provide robust estimates. This paper investigates *maximum a posteriori* (MAP) estimation of continuous density hidden Markov models (CDHMM). The derivations given here can straight-forwardly be extended to the subcases of discrete density HMM and tied-mixture HMM. The MAP estimate can be seen as a Bayes estimate of the vector parameter when the loss function is not specified [5]. The MAP estimation framework provides a way of incorporating prior information in the training process, which is particularly useful for dealing with problems posed by sparse training data for which the ML approach gives inaccurate estimates. MAP estimation can be applied to two classes of applications, namely, parameter smoothing and model adaptation, both related to the problem of parameter estimation with sparse training data.

In the following the sample $\mathbf{x} = (x_1, ..., x_T)$ denotes a given set of T observation vectors, where $x_1, ..., x_T$ are either independent and identically distributed (i.i.d.), or are drawn from a probabilistic function of a Markov chain.

[1] This work was done while Jean-Luc Gauvain was on leave from the Speech Communication Group at LIMSI/CNRS, Orsay, France.

The difference between MAP and ML estimation lies in the assumption of an appropriate prior distribution of the parameters to be estimated. If θ, assumed to be a random vector taking values in the space Θ, is the parameter vector to be estimated from the sample $\mathbf{x}$ with probability density function (p.d.f.) $f(\cdot|\theta)$, and if g is the prior p.d.f. of θ, then the MAP estimate, θ_{MAP}, is defined as the mode of the posterior p.d.f. of θ denoted as $g(\cdot|\mathbf{x})$, i.e.

$$\begin{aligned} \theta_{\mathrm{MAP}} &= \operatorname*{argmax}_{\theta} g(\theta|\mathbf{x}) &(1)\\ &= \operatorname*{argmax}_{\theta} f(\mathbf{x}|\theta)g(\theta). &(2)\end{aligned}$$

If θ is assumed to be fixed but unknown, then there is no knowledge about θ, which is equivalent to assuming a non-informative prior or an improper prior, i.e. $g(\theta)$ =constant. Under such an assumption, equation (2) then reduces to the familiar ML formulation.

Given the MAP formulation three key issues remain to be addressed: the choice of the prior distribution family, the specification of the parameters for the prior densities and the evaluation of the maximum a posteriori. These problems are closely related, since an appropriate choice of the prior distribution can greatly simplify the MAP estimation process.

Similar to ML estimation, MAP estimation is relatively easy if the family of p.d.f.'s $\{f(\cdot|\theta), \theta \in \Theta\}$ possesses a *sufficient statistic* of fixed dimension $t(\mathbf{x})$ for the parameter θ, i.e. $f(\mathbf{x}|\theta)$ can be factored into two terms $f(\mathbf{x}|\theta) = h(\mathbf{x})k(\theta|t(\mathbf{x}))$ such that $h(\mathbf{x})$ is independent of θ and $k(\theta|t(\mathbf{x}))$ is the *kernel density* which is a function of θ and depends on $\mathbf{x}$ only through the sufficient statistic $t(\mathbf{x})$ [27, 5, 7]. In this case, the natural solution is to choose the prior density in a *conjugate family* $\{k(\cdot|\varphi), \varphi \in \phi\}$, which includes the kernel density of $f(\cdot|\theta)$. The MAP estimation is then reduced to the evaluation of the mode of the posteriori density $k(\theta|\varphi') \propto k(\theta|\varphi)k(\theta|t(\mathbf{x}))$, a problem almost identical to the ML estimation problem of finding the mode of the kernel density $k(\cdot|t(\mathbf{x}))$. However, among the distribution families of interest, only exponential families have a sufficient statistic of fixed dimension [4, 17].

When there is no sufficient statistic of a fixed dimension, MAP estimation, like ML estimation, is a much more difficult problem because the posterior density is not expressible in terms of a fixed number of parameters and cannot be maximized easily. For both finite mixture densities and hidden Markov models, the lack of a sufficient statistic of a fixed dimension is due to the underlying hidden process, i.e. the state mixture component and the state sequence of a Markov chain for an HMM. In these cases ML estimates are usually obtained using the *expectation-maximization* (EM) algorithm [6, 1, 28]. For HMM parameter estimation this algorithm is also called the Baum-Welch algorithm. The EM algorithm is an iterative procedure for approximating ML estimates in the general case of models involving *incomplete data*. It locally maximizes the likelihood function of the observed (or incomplete) data. This algorithm exploits the fact that the complete-data likelihood is simpler to maximize than the likelihood of the incomplete data, as in the case where the *complete-data* model has sufficient statistics of fixed dimension. As noted by Dempster et al. [6], the EM algorithm can also be applied to MAP estimation.

The remainder of this paper is organized as follows. For HMM estimation, two types of random parameters are commonly used: one involves parameters that follow multinomial densities and the other involves parameters of multivariate Gaussian densities. In Section 2, the choice of the prior density family is addressed and it is shown that the prior densities for the HMM parameters can be adequately represented as a product of Dirichlet densities and normal-Wishart densities. Sections 3 and 4 derive formulations for MAP estimation of multivariate mixture Gaussian densities and for CDHMM with mixture Gaussian state observation densities. In Section 5, the important issue of prior density estimation is discussed. Some experimental results illustrating the practical interest of this approach are given in Section 6, and Bayesian Learning is shown to be a unified approach for a variety of applications including parameter smoothing and model adaptation. Finally our findings are summarized in Section 7.

2 Choices of Prior Densities

In this section the choice of the prior density family is addressed. Let $\mathbf{x} = (x_1, ..., x_T)$ be a sample of T i.i.d. observations drawn from a mixture of K p-dimensional multivariate normal densities. The joint p.d.f. is specified by the equation[2]

$$f(\mathbf{x}|\theta) = \prod_{t=1}^{T} \sum_{k=1}^{K} \omega_k \mathcal{N}(x_t|m_k, r_k) \tag{3}$$

where

$$\theta = (\omega_1, ..., \omega_K, m_1, ..., m_K, r_1, ..., r_K) \tag{4}$$

is the parameter vector and ω_k denotes the mixture gain for the k-th mixture component subject to the constraint $\sum_{k=1}^{K} \omega_k = 1$. $\mathcal{N}(x|m_k, r_k)$ is the k-th normal density function denoted by

$$\mathcal{N}(x|m_k, r_k) \propto |r_k|^{1/2} \exp[-\frac{1}{2}(x - m_k)^t r_k (x - m_k)] \tag{5}$$

where m_k is the p-dimensional mean vector and r_k is the $p \times p$ precision matrix[3].

As stated in the Introduction, no sufficient statistic of a fixed dimension exists for the parameter vector θ in equation (4), therefore no joint *conjugate prior density* can be specified. However a finite mixture density can be interpreted as a density associated with a statistical population which is a mixture of K component populations with mixing proportions $(\omega_1, \ldots, \omega_K)$. In other words, $f(\mathbf{x}|\theta)$ can be seen as a marginal p.d.f. of the joint p.d.f. of the parameter θ expressed as the product of a multinomial density (for the sizes of the component populations) and multivariate Gaussian densities (for the component densities). Consider that the mixture gains for each mixture density have the joint distribution is in the form of a multinomial distribution. Then a practical candidate to model the prior knowledge about the mixture gain parameter vector is the conjugate density such as a Dirichlet density [14]

$$g(\omega_1, ..., \omega_K|\nu_1, ..., \nu_K) \propto \prod_{k=1}^{K} \omega_k^{\nu_k - 1} \tag{6}$$

[2] In the following the same term f is used to denote both the joint and the marginal p.d.f.'s since it is not likely to cause confusion.

[3] $|r|$ denotes the determinant of the matrix r and r^t denotes the transpose of the matrix or vector r. In the following, we will also use $\mathrm{tr}(r)$ to denote the trace of the matrix r. A precision matrix is defined as the inverse of the covariance matrix.

where $\nu_k > 0$ are the parameters for the Dirichlet density. As for the vector parameter (m_k, r_k) of the individual Gaussian mixture component, the joint conjugate prior density is a normal-Wishart density [5] of the form

$$g(m_k, r_k|\tau_k, \mu_k, \alpha_k, u_k) \propto |r_k|^{(\alpha_k - p)/2} \exp[-\frac{\tau_k}{2}(m_k - \mu_k)^t r_k (m_k - \mu_k)] \exp[-\frac{1}{2}\mathrm{tr}(u_k r_k)] \quad (7)$$

where $(\tau_k, \mu_k, \alpha_k, u_k)$ are the prior density parameters such that $\alpha_k > p - 1$, $\tau_k > 0$, μ_k is a vector of dimension p and u_k is a $p \times p$ positive definite matrix.

Assuming independence between the parameters of the individual mixture components and the set of the mixture weights, the joint prior density $g(\theta)$ is the product of the prior p.d.f.'s defined in equations (6) and (7), i.e.

$$g(\theta) = g(\omega_1, ..., \omega_K) \prod_{k=1}^{K} g(m_k, r_k). \quad (8)$$

It will be shown that this choice for the prior density family can also be justified by noting that the EM algorithm can be applied to the MAP estimation problem if the prior density belongs to the conjugate family of the complete-data density.

3 MAP Estimates for Gaussian Mixture

The EM algorithm is an iterative procedure for approximating ML estimates in the context of incomplete-data cases such as mixture density and hidden Markov model estimation problems [2, 6, 28]. This procedure consists of maximizing at each iteration the auxiliary function $Q(\theta, \hat{\theta})$ defined as the expectation of the *complete-data* log-likelihood $\log h(\mathbf{y}|\theta)$ given the incomplete data $\mathbf{x} = (x_1, ..., x_T)$ and the current fit $\hat{\theta}$, i.e.

$$Q(\theta, \hat{\theta}) = E[\log h(\mathbf{y}|\theta)|\mathbf{x}, \hat{\theta}]. \quad (9)$$

For a mixture density, the complete-data likelihood is the joint likelihood of $\mathbf{x}$ and $\ell = (\ell_1, ..., \ell_T)$ the unobserved labels referring to the mixture components, i.e. $\mathbf{y} = (\mathbf{x}, \ell)$.

The EM procedure derives from the facts that $\log f(\mathbf{x}|\theta) = Q(\theta, \hat{\theta}) - H(\theta, \hat{\theta})$ where $H(\theta, \hat{\theta}) = E[\log h(\mathbf{y}|\mathbf{x}, \theta)|\mathbf{x}, \hat{\theta})]$ and $H(\theta, \hat{\theta}) \leq H(\hat{\theta}, \hat{\theta})$, and therefore whenever a value θ satisfies $Q(\theta, \hat{\theta}) > Q(\hat{\theta}, \hat{\theta})$ then $f(\mathbf{x}|\theta) > f(\mathbf{x}|\hat{\theta})$. It follows that the same iterative procedure can be used to estimate the mode of the posterior density by maximizing the auxiliary function $R(\theta, \hat{\theta}) = Q(\theta, \hat{\theta}) + \log g(\theta)$ at each iteration instead of the maximization of $Q(\theta, \hat{\theta})$ in conventional ML procedures [6].

For a mixture of K densities $\{f(\cdot|\theta_k)\}_{k=1,...,K}$ with mixture weights $\{\omega_k\}_{k=1,...,K}$, the auxiliary function Q takes the following form [28]:

$$Q(\theta, \hat{\theta}) = \sum_{t=1}^{T} \sum_{k=1}^{K} \frac{\hat{\omega}_k f(x_t|\hat{\theta}_k)}{\sum_{l=1}^{K} \hat{\omega}_l f(x_t|\hat{\theta}_l)} \log \omega_k f(x_t|\theta_k). \quad (10)$$

Let $\Psi(\theta, \hat{\theta}) = \exp R(\theta, \hat{\theta})$ be the function to be maximized. For the case of Gaussian mixture component, we have $f(x_t|\hat{\theta}_k) = \mathcal{N}(x_t|\hat{m}_k, \hat{r}_k)$. Define the following notations

$$\begin{aligned} c_{kt} &= \frac{\hat{\omega}_k \mathcal{N}(x_t|\hat{m}_k, \hat{r}_k)}{\sum_{l=1}^{K} \hat{\omega}_l \mathcal{N}(x_t|\hat{m}_l, \hat{r}_l)} && (11) \\ c_k &= \sum_{t=1}^{T} c_{kt} && (12) \\ \bar{x}_k &= \sum_{t=1}^{T} c_{kt} x_t / c_k && (13) \\ S_k &= \sum_{t=1}^{T} c_{kt}(x_t - \bar{x}_k)(x_t - \bar{x}_k)^t. && (14) \end{aligned}$$

Using the equality $\sum_{t=1}^{T} c_{kt}(x_t - m_k)^t r_k (x_t - m_k) = c_k (m_k - \bar{x}_k)^t r_k (m_k - \bar{x}_k) + \mathrm{tr}(S_k r_k)$, it follows from the definition of $f(\mathbf{x}|\theta)$ and equation (10) that

$$\Psi(\theta, \hat{\theta}) \propto g(\theta) \prod_{k=1}^{K} \omega_k^{c_k} |r_k|^{c_k/2} \exp[-\frac{c_k}{2}(m_k - \bar{x}_k)^t r_k (m_k - \bar{x}_k) - \tfrac{1}{2}\mathrm{tr}(S_k r_k)]. \qquad (15)$$

From the relations (15) and (8) it can easily be verified that $\Psi(\cdot, \hat{\theta})$ belongs to the same distribution family as $g(\cdot)$, and has parameters $\{\nu'_k, \tau'_k, \mu'_k, \alpha'_k, u'_k\}_{k=1,\ldots,K}$ satisfying the following conditions:

$$\begin{aligned} \nu'_k &= \nu_k + c_k && (16) \\ \tau'_k &= \tau_k + c_k && (17) \\ \alpha'_k &= \alpha_k + c_k && (18) \\ \mu'_k &= \frac{\tau_k \mu_k + c_k \bar{x}_k}{\tau_k + c_k} && (19) \\ u'_k &= u_k + S_k + \frac{\tau_k c_k}{\tau_k + c_k}(\mu_k - \bar{x}_k)(\mu_k - \bar{x}_k)^t && (20) \end{aligned}$$

The family of densities defined by (8) is therefore a conjugate family for the complete-data density.

The mode of $\Psi(\cdot, \hat{\theta})$, denoted $(\tilde{\omega}_k, \tilde{m}_k, \tilde{r}_k)$, may be obtained from the modes of the Dirichlet and normal-Wishart densities: $\tilde{\omega}_k = (\nu'_k - 1)/\sum_{k=1}^{K}(\nu'_k - 1)$, $\tilde{m}_k = \mu'_k$, and $\tilde{r}_k = (\alpha'_k - p){u'_k}^{-1}$. Thus, the EM reestimation formulas are derived as follows:

$$\begin{aligned} \tilde{\omega}_k &= \frac{(\nu_k - 1) + \sum_{t=1}^{T} c_{kt}}{\sum_{k=1}^{K}(\nu_k - 1) + \sum_{k=1}^{K}\sum_{t=1}^{T} c_{kt}} && (21) \\ \tilde{m}_k &= \frac{\tau_k \mu_k + \sum_{t=1}^{T} c_{kt} x_t}{\tau_k + \sum_{t=1}^{T} c_{kt}} && (22) \\ \tilde{r}_k^{-1} &= \frac{u_k + \sum_{t=1}^{T} c_{kt}(x_t - \tilde{m}_k)(x_t - \tilde{m}_k)^t + \tau_k(\mu_k - \tilde{m}_k)(\mu_k - \tilde{m}_k)^t}{(\alpha_k - p) + \sum_{t=1}^{T} c_{kt}}. && (23) \end{aligned}$$

For the Gaussian mean vectors, it can be seen that the new parameter estimates are simply a weighted sum of the prior parameters and the observed data. The above development suggests when the EM algorithm can be used for maximum likelihood estimation, a natural prior density can be found in the conjugate family of the complete-data density if such a conjugate family exists. For example, in the general case of mixture densities from exponential families, the prior will be the product of a Dirichlet density for the mixture weights and the conjugate densities of the mixture components.

If it is assumed that each mixture component is non-degenerate, i.e. $\hat{\omega}_k > 0$, then $c_{k1}, c_{k2}, ..., c_{kT}$ is a sequence of T i.i.d. random variables with a non-degenerate distribution and $\limsup_{T\to\infty} \sum_{t=1}^{T} c_{kt} = \infty$ with probability one [25]. It follows that $\tilde{w_k}$ converges to $\sum_{t=1}^{T} c_{kt}/T$ with probability one when $T \to \infty$. Applying the same reasoning to $\tilde{m_k}$ and $\tilde{r_k}$, it can be seen that the EM reestimation formulas for the MAP and ML approaches are asymptotically similar. Thus as long as the initial estimates of $\hat{\theta}$ are identical, the EM algorithms for MAP and ML will provide identical estimates with probability one when $T \to \infty$.

4 MAP Estimates for HMM

The development in the previous section for a mixture of multivariate Gaussian densities can be extended to the case of HMM with Gaussian mixture state observation densities. For notational convenience, it is assumed that the observation p.d.f.'s of all the states have the same number of mixture components.

Consider an N-state HMM with parameter vector $\lambda = (\pi, \mathbf{A}, \theta)$, where π is the initial probability vector, $\mathbf{A}$ is the transition matrix, and θ is the p.d.f. parameter vector composed of the mixture parameters $\theta_i = \{w_{ik}, m_{ik}, r_{ik}\}_{k=1,...,K}$ for each state i.

For a sample $\mathbf{x} = (x_1, ..., x_T)$, the complete data is $\mathbf{y} = (\mathbf{x}, \mathbf{s}, \ell)$ where $\mathbf{s} = (s_0, ..., s_T)$ is the unobserved state sequence, and $\ell = (\ell_1, ..., \ell_T)$ is the sequence of the unobserved mixture component labels, $s_t \in [1, N]$ and $l_t \in [1, K]$. The joint p.d.f. $h(\cdot|\lambda)$ of **x**, **s**, and ℓ is defined as[4]

$$h(\mathbf{x}, \mathbf{s}, \ell|\lambda) = \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} \omega_{s_t\ell_t} f(x_t|\theta_{s_t\ell_t}) \tag{24}$$

where π_i is the initial probability of state i, a_{ij} is the transition probability from state i to state j, and $\theta_{ik} = (m_{ik}, r_{ik})$ is the parameter vector of the k-th normal p.d.f. associated with state i. It follows that the likelihood of **x** has the form

$$f(\mathbf{x}|\lambda) = \sum_{\mathbf{s}} \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} f(x_t|\theta_{s_t}) \tag{25}$$

where $f(x_t|\theta_i) = \sum_{k=1}^{K} \omega_{ik} \mathcal{N}(x_t|m_{ik}, r_{ik})$, and the summation is over all possible state sequences.

If no prior knowledge is assumed about **A** and π, or alternatively if these parameters are assumed fixed and known, the prior density G can be chosen to have the following form $G(\lambda) = \prod_i g(\theta_i)$, where $g(\theta_i)$ is defined by equation (8). In the general case where MAP estimation is applied not only to the observation density parameters but also to the initial and transition probabilities, a Dirichlet density can be used for the initial probability vector π and for each row of the transition probability matrix **A**. This

[4] Here we use the definition proposed by Baum et al. [1], where the observation p.d.f.'s are associated to the Markov chain states and no symbol is produced in state s_0

choice follows directly from the derivation discussed in the previous section, since the complete-data likelihood satisfies $h(\mathbf{x}, \mathbf{s}, \ell|\lambda) = h(\mathbf{s}|\lambda)h(\mathbf{x}, \ell|\mathbf{s}, \lambda)$ where $h(\mathbf{s}|\lambda)$ is the product of $N+1$ multinomial densities with parameter sets $\{\pi_1, ..., \pi_N\}$ and $\{a_{i1}, ..., a_{iN}\}_{i=1,...,N}$. The prior density for all the HMM parameters thus satisfies the relation

$$G(\lambda) \propto \prod_{i=1}^{N} \left[\pi_i^{\eta_i - 1} g(\theta_i) \prod_{j=1}^{N} a_{ij}^{\eta_{ij}-1} \right] \tag{26}$$

where $\{\eta_i\}$ is the set of parameters for the prior density of the initial probabilities $\{\pi_i\}$, and $\{\eta_{ij}\}$ is the set of parameters for the prior density of transition probabilities $\{a_{ij}\}$ defined the same way as in equation (6).

In the following subsections we examine two ways of approximating λ_{MAP} by local maximization of $f(\mathbf{x}|\lambda)G(\lambda)$ or of $f(\mathbf{x}, \mathbf{s}|\lambda)G(\lambda)$. These two solutions are the MAP versions of the Baum-Welch algorithm [2] and of the segmental k-means algorithm [26], algorithms which were developed for ML estimation.

4.1 Forward-Backward MAP Estimate

From equation (24) it is straightforward to show that the auxiliary function of the EM algorithm applied to ML estimation of λ, $Q(\lambda, \hat{\lambda}) = E[\log h(\mathbf{y}|\lambda)|\mathbf{x}, \hat{\lambda}]$, can be decomposed into a sum of three auxiliary functions: $Q_\pi(\pi, \hat{\lambda})$, $Q_A(\mathbf{A}, \hat{\lambda})$ and $Q_\theta(\theta, \hat{\lambda})$ such that they can be independently maximized [15]. The three functions take the following forms:

$$\begin{aligned}
Q_\pi(\pi, \hat{\lambda}) &= \sum_{i=1}^{N} \gamma_{i0} \log \pi_i & (27)\\
Q_A(\mathbf{A}, \hat{\lambda}) &= \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N} \Pr(s_{t-1} = i, s_t = j|\mathbf{x}, \hat{\lambda}) \log a_{ij} & (28)\\
&= \sum_{i=1}^{N} Q_{a_i}(a_i, \hat{\lambda}) & (29)\\
Q_\theta(\theta, \hat{\lambda}) &= \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{K} \Pr(s_t = i, \ell_t = k|\mathbf{x}, \hat{\lambda}) \log \omega_{ik} f(x_t|\theta_{ik}) & (30)\\
&= \sum_{i=1}^{N} Q_{\theta_i}(\theta_i|\hat{\lambda}) & (31)
\end{aligned}$$

with

$$\begin{aligned}
Q_{a_i}(a_i, \hat{\lambda}) &= \sum_{t=1}^{T}\sum_{j=1}^{N} \xi_{ijt} \log a_{ij} & (32)\\
Q_{\theta_i}(\theta_i, \hat{\lambda}) &= \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_{it} \frac{\hat{\omega}_{ik} f(x_t|\hat{\theta}_{ik})}{\sum_{l=1}^{K} \hat{\omega}_{il} f(x_t|\hat{\theta}_{il})} \log \omega_{ik} f(x_t|\theta_{ik}) & (33)
\end{aligned}$$

where $\xi_{ijt} = \Pr(s_{t-1} = i, s_t = j|\mathbf{x}, \hat{\lambda})$ is the probability of making a transition from state i to state j at time t given that the model $\hat{\lambda}$ generates **x**, and $\gamma_{it} = \Pr(s_t = i|\mathbf{x}, \hat{\lambda})$ is the probability of being in state i at time t given that the model $\hat{\lambda}$ generates **x**. Both probabilities can be computed at each EM iteration using the forward-backward algorithm [2].

As for the mixture Gaussian case, estimating the mode of the posterior density requires the maximization of the auxiliary function $R(\lambda, \hat{\lambda}) = Q(\lambda, \hat{\lambda}) + \log G(\lambda)$. The form chosen for $G(\lambda)$ in (26) permits independent maximization of each of the following $2N + 1$ parameter sets: $\{\pi_1, ..., \pi_N\}$, $\{a_{i1}, ..., a_{iN}\}_{i=1,...,N}$ and $\{\theta_i\}_{i=1,...,N}$. The MAP auxiliary function $R(\lambda, \hat{\lambda})$ can thus be written as the sum $R_\pi(\pi, \hat{\lambda}) + \sum_i R_{a_i}(a_i, \hat{\lambda}) + \sum_i R_{\theta_i}(\theta_i, \hat{\lambda})$, where each term represents the MAP auxiliary function associated with the respective indexed parameter sets.

We can recognize in (33) the same form as was seen for $Q(\theta, \hat{\theta})$ in (10) for the mixture Gaussian case. It follows that if c_{kt} in equation (11) is replaced by c_{ikt} defined as

$$c_{ikt} = \gamma_{it} \frac{\hat{\omega}_{ik} \mathcal{N}(x_t|\hat{m}_{ik}, \hat{r}_{ik})}{\sum_{l=1}^{K} \mathcal{N}(x_t|\hat{m}_{il}, \hat{r}_{il})} \tag{34}$$

which is the probability of being in state i with the mixture component label k at time t given that the model $\hat{\lambda}$ generates x_t, then the reestimation formulas (21-23) can be used to maximize $R_{\theta_i}(\theta_i, \hat{\lambda})$.

It is straightforward to derive the reestimations formulas for π and **A** by applying the same derivations as were used for the mixture weights. The EM iteration for the three parameter set $\lambda = (\pi, \mathbf{A}, \theta)$ is:

$$\tilde{\pi}_i = \frac{(\eta_i - 1) + \gamma_{i0}}{\sum_{j=1}^{N} (\eta_j - 1) + \sum_{j=1}^{N} \gamma_{j0}} \tag{35}$$

$$\tilde{a}_{ij} = \frac{(\eta_{ij} - 1) + \sum_{t=1}^{T} \xi_{ijt}}{\sum_{j=1}^{N} (\eta_{ij} - 1) + \sum_{j=1}^{N} \sum_{t=1}^{T} \xi_{ijt}} \tag{36}$$

$$\tilde{\omega}_{ik} = \frac{(\nu_{ik} - 1) + \sum_{t=1}^{T} c_{ikt}}{\sum_{k=1}^{K} (\nu_{ik} - 1) + \sum_{k=1}^{K} \sum_{t=1}^{T} c_{ikt}} \tag{37}$$

$$\tilde{m}_{ik} = \frac{\tau_{ik} \mu_{ik} + \sum_{t=1}^{T} c_{ikt} x_t}{\tau_{ik} + \sum_{t=1}^{T} c_{ikt}} \tag{38}$$

$$\tilde{r}_{ik}^{-1} = \frac{u_{ik} + \sum_{t=1}^{T} c_{ikt} (x_t - \tilde{m_{ik}})(x_t - \tilde{m_{ik}})^t + \tau_{ik} (\mu_{ik} - \tilde{m_{ik}})(\mu_{ik} - \tilde{m_{ik}})^t}{(\alpha_{ik} - p) + \sum_{t=1}^{T} c_{ikt}}. \tag{39}$$

For multiple independent observation sequences $\{\mathbf{x}_v\}_{v=1,...,V}$, with $\mathbf{x}_v = (x_1^{(v)}, ..., x_{T_v}^{(v)})$, we must maximize $G(\lambda) \prod_{v=1}^{V} f(\mathbf{x}_v|\lambda)$, where $f(\cdot|\lambda)$ is defined by equation (25). The EM auxiliary function is then $R(\lambda, \hat{\lambda}) = \log G(\lambda) + \sum_{v=1}^{V} E[\log h(\mathbf{y}_v|\lambda)|\mathbf{x}_v, \hat{\lambda}]$, where $h(\cdot|\lambda)$ is defined by equation (24). It follows that the reestimation formulas for **A** and θ still hold if the summations over t ($\sum_{t=1}^{T}$) are replaced by summations over v and t ($\sum_{v=1}^{V} \sum_{t=1}^{T_v}$). The values $\xi_{ijt}^{(v)}$ and $\gamma_{it}^{(v)}$ are then obtained by applying the forward-backward algorithm for each observation sequence. The reestimation formula for the initial probabilities becomes

$$\tilde{\pi}_i = \frac{(\eta_i - 1) + \sum_{v=1}^{V} \gamma_{i0}^{(v)}}{\sum_{j=1}^{N} (\eta_j - 1) + \sum_{j=1}^{N} \sum_{v=1}^{V} \gamma_{j0}^{(v)}}. \tag{40}$$

Reestimation formulation similar to equations (36-39) can also be derived. Just like for the mixture parameter case, it can be shown that as $V \to \infty$, the MAP reestimation formulas approach the ML ones, exhibiting the asymptotical similarity of the two estimates.

These reestimation equations give estimates of the HMM parameters which correspond to a local maximum of the posterior density. The choice of the initial estimates is therefore critical to ensure a solution close to the global maximum and to minimize the number of EM iterations needed to attain the local maximum. When using an informative prior, a natural choice for the initial estimates is the mode of the prior density, which represents all the available information about the parameters when no data has been observed. The corresponding values are simply obtained by applying the reestimation formulas with T equal to 0 (i.e. without any observed data). Unlike the case of discrete HMMs where it is possible to use uniform initial estimates, there is no trivial initial solution for the continuous density HMM case. Therefore, in practice, the statistician adds information in the training process such as a uniform or manual segmentation of the observation sequence into states from which it is possible to obtain raw estimates of the HMM parameters by direct computation of the mode of the complete-data likelihood.

4.2 Segmental MAP Estimate

By analogy with the segmental k-means algorithm [26], a similar optimization criterion can be adopted. Instead of maximizing $G(\lambda|\mathbf{x})$, the joint posterior density of parameter λ and state sequence **s**, $G(\lambda, \mathbf{s}|\mathbf{x})$, is maximized. The estimation procedure becomes

$$\tilde{\lambda} = \operatorname*{argmax}_{\lambda} \max_{\mathbf{s}} G(\lambda, \mathbf{s}|\mathbf{x}) = \operatorname*{argmax}_{\lambda} \max_{\mathbf{s}} f(\mathbf{x}, \mathbf{s}|\lambda) G(\lambda). \tag{41}$$

where $\tilde{\lambda}$ is refered to as the *segmental MAP estimate* of λ. As for the segmental k-means algorithm [16], it is straightforward to prove that starting with any estimate $\lambda^{(m)}$, alternate maximization over **s** and λ gives a sequence of estimates with non-decreasing values of $G(\lambda, \mathbf{s}|\mathbf{x})$, i.e. $G(\lambda^{(m+1)}, \mathbf{s}^{(m+1)}|\mathbf{x}) \geq G(\lambda^{(m)}, \mathbf{s}^{(m)}|\mathbf{x})$ with

$$\begin{aligned} \mathbf{s}^{(m+1)} &= \operatorname*{argmax}_{\mathbf{s}} f(\mathbf{x}, \mathbf{s}|\lambda^{(m)}) && (42) \\ \lambda^{(m+1)} &= \operatorname*{argmax}_{\lambda} f(\mathbf{x}, \mathbf{s}^{(m+1)}|\lambda) G(\lambda). && (43) \end{aligned}$$

The most likely state sequence $\mathbf{s}^{(m+1)}$ is decoded by the Viterbi algorithm [9]. Maximization over λ can also be replaced by any *hill climbing* procedure over λ subject to the constraint that $f(\mathbf{x}, \mathbf{s}^{(m+1)}|\lambda^{(m+1)})G(\lambda^{(m+1)}) \geq f(\mathbf{x}, \mathbf{s}^{(m+1)}|\lambda^{(m)})G(\lambda^{(m)})$. The EM algorithm is once again a good candidate to perform this maximization using $\lambda^{(m)}$ as an initial estimate. The EM auxiliary function is then $R(\lambda, \hat{\lambda}) = \log G(\lambda) + E[\log h(\mathbf{y}|\lambda)|\mathbf{x}, \mathbf{s}^{(m)}, \hat{\lambda}]$ where $h(\cdot|\lambda)$ is defined by equation (24). It is straightforward to show that the reestimation equations (35-39) still hold with $\xi_{ijt} = \delta(s_{t-1}^{(m)} - i)\delta(s_t^{(m)} - j)$ and $\gamma_{it} = \delta(s_t^{(m)} - i)$, where δ denotes the Kronecker delta function.

5 Prior density estimation

In the previous sections it was assumed that the prior density $G(\lambda)$ is a member of a preassigned family of prior distributions defined by (26). In a strictly Bayesian approach the vector parameter φ of this family of p.d.f.'s $\{G(\cdot|\varphi), \varphi \in \phi\}$ is also assumed known based on common or subjective knowledge about the stochastic process. An alternate solution is to adopt an empirical Bayes approach [29] where the prior parameters are estimated directly from data. The estimation is then based on the marginal distribution of the data given the estimated prior parameters.

In fact, part of the available prior knowledge can be directly incorporated in the model by assuming some of the parameters to be fixed and known and/or by tying some of the parameters. As for the prior distribution, this information will reduce the uncertainty during the training process and increase the robustness of the estimates. However, in contrast to the prior distribution, such deterministic prior information by definition cannot be changed even if a large amount of training data is available.

Adopting the empirical Bayes approach, it is assumed that the sequence of observations, **X**, is composed of multiple independent sequences associated with different unknown values of the HMM parameters. Let $(\mathbf{X}, \Lambda) = [(\mathbf{x}_1, \lambda_1), (\mathbf{x}_2, \lambda_2), ..., (\mathbf{x}_Q, \lambda_Q)]$ be such a multiple sequence of observations, where each pair is independent of the others and the λ_q have a common prior distribution $G(\cdot|\varphi)$. Since the λ_q are not directly observed, the prior parameter estimates must be obtained from the marginal density $f(\mathbf{X}|\varphi)$ defined as

$$f(\mathbf{X}|\varphi) = \int_\Lambda f(\mathbf{X}|\Lambda)G(\Lambda|\varphi)\, d\Lambda \tag{44}$$

where $f(\mathbf{X}|\Lambda) = \prod_q f(\mathbf{x}_q|\lambda_q)$ and $G(\Lambda|\varphi) = \prod_q G(\lambda_q|\varphi)$. However, maximum likelihood estimation based on $f(\mathbf{X}|\varphi)$ appears rather difficult. To alleviate the problem, we can choose a simpler optimization criterion of maximizing the joint p.d.f. $f(\mathbf{X}, \Lambda|\varphi)$ over Λ and φ instead of maximizing the marginal p.d.f. of **X** given φ. Starting with an initial estimate of $\varphi^{(m)}$, a hill climbing procedure is obtained by alternate maximization over Λ and φ, i.e.

$$\Lambda^{(m)} = \operatorname*{argmax}_\Lambda f(\mathbf{X}, \Lambda|\varphi^{(m)}) \tag{45}$$

$$\varphi^{(m+1)} = \operatorname*{argmax}_\varphi G(\Lambda^{(m)}|\varphi) \tag{46}$$

Such a procedure provides a sequence of estimates with non-decreasing values of $f(\mathbf{X}, \Lambda|\varphi^{(m)})$. The solution of (45) is the MAP estimate of Λ based on the current prior parameter $\varphi^{(m)}$, which can be obtained by applying the forward-backward MAP reestimation formulas to each observation sequence $\mathbf{x}_q$. The solution of (46) is the maximum likelihood estimate of φ based on the current values of the HMM parameters. It should be noted that this procedure gives not only an estimate of the prior parameters but also MAP estimates of the HMM parameters for each independent observation sequence $\mathbf{x}_q$.

Finding the solution of equation (46) poses two problems. First, due to the Wishart and Dirichlet components, maximum likelihood estimation for the density defined by (26) is not trivial. Second, since more parameters are needed for the prior density than for the HMM itself, there can be a problem of overparametrization when the number of pairs $(\mathbf{x}_q, \lambda_q)$ is small. One way to simplify the estimation problem is to use moment estimates to approximate the ML estimates. For the overparametrization problem, it is possible to reduce the size of the prior family by adding some constraints on the prior parameters. For example, the prior family can be limited to the family of the kernel density of the complete-data likelihood, i.e. the posterior density family of the complete-data model when no prior information is available. Doing so, it is easy to show that the following constraints on the prior parameters hold

$$\nu_{ik} = \tau_{ik} \tag{47}$$

$$\alpha_{ik} = \tau_{ik} + p. \tag{48}$$

Parameter tying can also be used to further reduce the size of the prior family and is useful for parameter smoothing purposes. Finally, another practical constraint is to impose the prior mode to be equal to the parameters of a given HMM, resulting in a scheme for model adaptation.

This approach can be used for two classes of applications: parameter smoothing and adaptive learning. For parameter smoothing, the goal is to estimate $\{\lambda_1, \lambda_2, ..., \lambda_Q\}$. The abovementioned algorithm offers a direct solution to "smooth" these different estimates by assuming a common prior density for all the models. For adaptive learning, we observe a new sequence of observations $\mathbf{x}_q$ associated with the unobserved vector parameter value λ_q. The required specification of the prior parameters for finding the MAP estimate of λ_q can be obtained as a point estimate $\hat{\varphi}$ computed with the proposed iterative algorithm. Such a training process can be seen as the adaptation of a less specific a priori model $\hat{\lambda} = \mathrm{argmax}_\lambda G(\lambda|\hat{\varphi})$ (when no training data are available) to more specific conditions which match well with the new observation sequence $\mathbf{x}_q$. Some experimental results for these applications are given in the next section.

6 Experimental Results

Bayesian learning of Gaussian densities has been widely used for sequential learning of the mean vectors of feature- and template-based recognizers (see for example, Zelinski and Class [31], Stern and Lasry [30]). Ferretti and Scarci [8] used Bayesian estimation of mean vectors to build speaker-specific codebooks in an HMM framework. In all these cases, the precision parameter was assumed to be known and the prior density limited to a Gaussian. Brown *et al.* [3] used Bayesian estimation for speaker adaptation of CDHMM parameters in a connected digit recognizer. More recently, Lee *et al.* [20] investigated various training schemes of Gaussian mean and variance parameters using normal-gamma prior densities for speaker adaptation. They showed that on the alpha-digit vocabulary, with only a small amount of speaker specific data (1 to 3 utterances of each word), the MAP estimates gave better results than the ML estimates.

Using the theoretical developments presented in this paper, Bayesian estimation has been successfully applied to CDHMM with Gaussian mixture observation densities for four speech recognition applications: parameter smoothing, speaker adaptation, speaker group modeling and corrective training. We have previously reported experimental results for these applications in [10, 11, 12, 22]. In order to demonstrate the effectiveness of Bayesian estimation for such applications, some results are given here. In all cases, the HMM parameters were estimated using the segmental MAP algorithm. The prior parameters, subject to the conditions (47-48), were obtained by forcing the prior mode to be equal to the parameters of a given HMM [10]. These constraints leave free the parameters τ_{ik} which can either be estimated using the algorithm described in Section 5, or can be arbitrarily fixed. For model adaptation, τ_{ik} can be regarded as a weight associated with the k-th Gaussian of state i as shown in equations (35) and (39). When this weight is large, the prior density is sharply peaked around the values of the seed HMM parameters which are only slightly modified by the adaptation process. Conversely, if τ_{ik} is small, adaptation is fast and the MAP estimates depend mainly on the observed data.

The applications discussed here are parameter smoothing and speaker adaptation. It is well known that HMM training requires smoothing (or tying), particularly if a large number of context-dependent (CD) phone models are used with limited amounts of training data. While several solutions have been investigated to smooth discrete HMMs, such as model interpolation, co-occurrence smoothing, and fuzzy VQ, only variance smoothing has been proposed for continuous density HMMs. In [10, 11] we have shown that MAP estimation can be used to solve this problem for CDHMMs by tying the parameters of the prior density. Performance improvement has been reported by tying the prior parameters in two ways. For CD model smoothing, the same prior density was used for all CD models corresponding to the same phone [10], and for p.d.f. smoothing the same marginal prior density was used for all the components of a given mixture [11]. In experiments using the DARPA Naval Resource Management (RM) [24] and the TI connected digit corpora, MAP estimation always outperformed ML estimation, with error rate reductions on the order of 10 to 25%.

In the case of model adaptation, MAP estimation may be viewed as a process for adjusting seed models to form more specific ones based on a small amount of adaptation data. The seed models are used to estimate the parameters of the prior densities and to serve as an initial estimate for the EM algorithm. Here experimental results are presented on speaker-adaptation as an example of model adaptation (Bayesian learning was also demonstrated as a scheme for sex-dependent training in [10, 11, 12].) The experiments used a set of context-independent (CI) phone models, where each model is a left-to-right HMM with Gaussian mixture state observation densities, with a maximum of 32 mixture components per state. Diagonal covariance matrices are used and the transition probabilities are assumed fixed and known. Details of the recognition system and the basic assumptions for acoustic modeling of subword units can be found in [19]. As described in [21], a 38-dimensional feature vector composed of LPC-derived cepstrum coefficients, and first and second order time derivatives was computed after the data were down-sampled to 8kHz to simulate the telephone bandwidth.

In Table 1, speaker adaptation using MAP estimation is compared to ML training of speaker-dependent (SD) models, using a set of 47 CI phone models. For MAP estimation speaker-independent (SI) and sex-dependent (M/F) seed models were trained on the standard RM SI-109 training set consisting of 3990 utterances from 109 native American talkers (31 females and 78 males), each providing 30 or 40 utterances. The test material consisted of the RM FEB91-SD test data with 25 testing utterances from each of the 12 testing speakers (7 males and 5 females). Results are reported using 40, 100 and 600 utterances (or equivalently about two, five and thirty minutes of speech material) of the speaker-specific data (taken from RM SD data) for training and adaptation. The MLE (SD) and MAP (SI) word

Training	0 min	2 min	5 min	30 min
MLE	—	31.5	12.1	3.5
MAP (SI)	13.9	8.7	6.9	3.4
MAP (M/F)	11.5	7.5	6.0	3.5

Table 1: Summary of SD, SA (SI), and SA (M/F) results on FEB91-SD test. Results are given as word error rate (%).

error rates using the standard RM word pair grammar are given in the two first rows of the table. The MLE (SD) word error rate for 2 minutes of training data is 31.5%. The SI word error rate (0 minutes of adaptation data) is 13.9%, somewhat comparable to the MLE result with 5 minutes of speaker-specific training data. While the MAP models are seen to outperform MLE models when only relatively small amounts of data were used for training or adaptation, the MAP and MLE results are comparable when all the available training data were used. This result is consistent with the Bayesian formulation that the MAP estimate and the MLE are asymptotically similar as demonstrated in equations (35) - (39) with $T \to \infty$. Compared to the SI results, the word error reduction is 37% with 2 minutes of adaptation data. A larger improvement was observed for the female speakers (51%) than for the male speakers (22%), presumably because there are fewer female speakers in the SI-109 training data.

Speaker adaptation can also be done using sex-dependent seed models if the gender of the new speaker is known or can be estimated prior to the adaptation process. In the case of estimation, the gender-dependent model set that best matches the gender of the new speaker is then used as the seed model set instead of the SI seed models. Results for speaker adaptation using sex-dependent seed models are given in the third row of Table 1. The word error rate without speaker adaptation is 11.5%. The error rate is reduced to 7.5% with 2 minutes, and 6.0% with 5 minutes, of adaptation data. Comparing the last 2 rows of the table it can be seen that speaker adaptation is more effective when sex-dependent seed models are used. The error reduction with 2 minutes of training data is 35% compared to the sex-dependent model results and 46% compared to the SI model results.

More details on experimental results using MAP estimation for parameter smoothing and model adaptation can be found in [10, 11, 12, 22] including application to speaker clustering and corrective training. MAP estimation has also been applied to task adaptation[22]. In this case task-independent SI models, trained from 10,000 utterance of a general English corpus[13], served as seed models for speaker and task adaptation. Another use of MAP estimation has recently been proposed for text-independent speaker identification[18] using a small amount of speaker-specific training data.

7 Conclusion

The theoretical framework for MAP estimation of multivariate Gaussian mixture density and HMM with Gaussian mixture state observation densities was presented. By extending the two well-known ML estimation algorithms to MAP estimation, two corresponding MAP training algorithms, namely the *forward-backward MAP estimation* and the *segmental MAP estimation*, were formulated. The proposed Bayesian estimation approach provides a framework to solve various HMM estimation problems posed by sparse training data. It has been applied successfully to acoustic modeling in automatic speech recognition, where Bayesian learning serves as a unified approach for speaker adaptation, speaker group modeling, parameter smoothing and corrective training. The same framework can also be adopted for the smoothing and adaptation of discrete and tied-mixture hidden Markov models (also known as semi-continuous hidden Markov models) and N-gram stochastic language models.

References

[1] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.*, vol. 41, pp. 164-171, 1970

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

[3] P. Brown, C.-H. Lee and J. Spohrer, "Bayesian Adaptation in Speech Recognition," *Proc. ICASSP-83*, pp. 761-764, 1983.

[4] G. Darmois, "Sur les lois de probabilité à estimation exhaustive," *C. R. Acad. Sci.*, 260, pp. 1265-1266, 1935.

[5] M. DeGroot, *Optimal Statistical Decisions*, McGraw-Hill, 1970.

[6] A. Dempster, N. Laird and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM algorithm," *J. Roy. Statist. Soc. Ser. B*, 39, pp. 1-38, 1977.

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

[8] M. Ferretti and S. Scarci, "Large-Vocabulary Speech Recognition with Speaker-Adapted Codebook and HMM Parameters," *Proc. EuroSpeech-89*, pp. 154-156, 1989.

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

[10] J.-L. Gauvain and C.-H. Lee, "Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models," *Proc. DARPA Speech and Natural language Workshop*, Pacific Grove, February 1991.

[11] J.-L. Gauvain and C.-H. Lee, "MAP Estimation of Continuous Density HMM: Theory and Applications," *Proc. DARPA Speech and Natural language Workshop*, Arden House, February 1992.

[12] J.-L. Gauvain and C.-H. Lee, "Bayesian Learning for Hidden Markov Model with Gaussian Mixture State Observation Densities," *Speech Communication*, vol. 11, nos. 2-3, June 1992.

[13] H.-W. Hon, “Vocabulary-Independent Speech Recognition: The VOCIND System,” *Ph. D. Thesis*, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, March 1992.

[14] N. L. Johnson and S. Kotz, *Distribution in Statistics*, John Wiley & Sons, New York, 1972.

[15] B.-H. Juang, “Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains,” *AT&T Technical Journal*, vol. 64, no. 6, July-August 1985.

[16] 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. ASSP-38, no. 9, September 1990.

[17] B. O. Koopman, “On distributions admitting a sufficient statistic”, *Trans. Am. Math. Soc.*, vol. 39, pp. 399-409, 1936.

[18] L. F. Lamel and J.-L. Gauvain, “Cross-Lingual Experiments with Phone Recognition,” to appear in *Proc. IEEE ICASSP-93*.

[19] C.-H. Lee, L. R. Rabiner, R. Pieraccini and J. G. Wilpon, “Acoustic modeling for large vocabulary speech recognition,” *Computer Speech and Language* 4, pp. 127-165, 1990.

[20] C.-H. Lee, C.-H. Lin and B.-H. Juang, “A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-39, no. 4, pp. 806-814, April 1991.

[21] C.-H. Lee, E. Giachin, L. R. Rabiner, R. Pieraccini and A. E. Rosenberg, “Improved acoustic modeling for large vocabulary continuous speech recognition,” *Computer Speech and Language* 6, pp. 103-127, 1992.

[22] C.-H. Lee and J.-L. Gauvain, “Speaker Adaptation Based on MAP Estimation of HMM Parameters,” to appear in *Proc. IEEE ICASSP-93*.

[23] L. R. Liporace, “Maximum Likelihood Estimation for Multivariate Observations of Markov Sources,” *IEEE Trans. Inform. Theory*, vol. IT-28, no. 5, pp. 729-734, September 1982.

[24] P. J. Price, W. Fisher, J. Bernstein, and D. Pallett, “A Database for Continuous Speech Recognition in a 1000-Word Domain,” *Proc. ICASSP-88*, New York, pp. 651-654, April 1988.

[25] Y. V. Prohorov and Y. A. Rozanov, *Probability Theory*, Springer-Verlag, 1969.

[26] 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. 64, no. 3, pp. 21-40, May 1986.

[27] C. R. Rao, *Linear Statistical Inference and Its Applications*, 2nd Edition, John Wiley & Sons, New York, 1973.

[28] R. A. Redner and H. F. Walker, “Mixture Densities, Maximum Likelihood and the EM Algorithm,” *SIAM Review*, vol. 26, no. 2, pp. 195-239, April 1984.

[29] H. Robbins, “The Empirical Bayes Approach to Statistical Decision Problems,” *Ann. Math. Statist.*, vol. 35, pp. 1-20, 1964.

[30] R. Stern and M. Lasry, “Dynamic Speaker Adaptation for Feature-Based Isolated Word Recognition,” *IEEE Trans. on ASSP*, Vol. ASSP-35, No. 6, June 1987.

[31] R. Zelinski and F. Class, “A learning procedure for speaker-dependent word recognition systems based on sequential processing of input tokens,” *Proc. ICASSP-83*, pp. 1053-1056, Boston, 1983.