Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains

Jean-Luc Gauvain and Chin-Hui Lee, *Senior Member, IEEE*

***Abstract*—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 HMM's 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.**

I. Introduction

ESTIMATION of a probabilistic function of Markov chain, which is 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's). The derivations given here can straightforwardly be extended to the subcases of discrete density HMM's and tied-mixture HMM's. 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, which are both related to the problem of parameter estimation with sparse training data.

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

Manuscript received June 18, 1992; revised June 20, 1993. The associate editor coordinating the review of this paper and approving it for publication was Dr. Xuedong Huang.

J. -L. Gauvain is with the Speech Communication Group at LIMSI/CNRS, Orsay, France.

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

IEEE Log Number 9215232.

The difference between MAP and ML estimation lies in the assumption of an appropriate prior distribution of the parameters to be estimated. If θ, which is 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.

$$\theta_{\mathrm{MAP}} = \operatorname*{argmax}_{\theta} g(\theta|\mathbf{x}) \tag{1}$$

$$= \operatorname*{argmax}_{\theta} f(\mathbf{x}|\theta)g(\theta). \tag{2}$$

If θ is assumed to be fixed but unknown, then there is no knowledge about θ, which is equivalent to assuming a noninformative prior or an improper prior, i.e., $g(\theta) =$ constant. Under such an assumption, (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 MAP. 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 $k(\theta|\varphi') \propto k(\theta|\varphi)k(\theta|t(\mathbf{x}))$, which is a problem that is 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 II, 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 III and IV derive formulations for MAP estimation of multivariate mixture Gaussian densities and for CDHMM with mixture Gaussian state observation densities. In Section V, the important issue of prior density estimation is discussed. Some experimental results illustrating the practical interest of this approach are given in Section VI, 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 VII.

II. Choices of Prior Densities

In this section, the choice of the prior density family is addressed. Let $\mathbf{x} = (x_1, \ldots x_T)$ 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[1]

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

where

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

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

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

where m_k is the p-dimensional mean vector, and r_k is the $p \times p$ precision matrix.[2]

As stated in the Introduction, no sufficient statistic of a fixed dimension exists for the parameter vector θ in (4), and 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 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). Assume that the joint density of the mixture gains for each mixture density is a multinomial distribution. Then, a practical candidate to model the prior knowledge about the mixture gain parameter vector is a conjugate density such as the Dirichlet density [14]

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

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

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

where $(\tau_k, \mu_k, \alpha_k, u_k)$ are the prior density parameters such that $\alpha_k > p - 1, \tau_k > 0, \mu_k$ 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 (6) and (7), i.e.

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

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.

III. 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 $\mathbf{x} = (x_1, \ldots, x_T)$ and the current fit $\hat{\theta}$, i.e.

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

For a mixture density, the complete-data likelihood is the joint likelihood of $\mathbf{x}$ and $\ell = (\ell_1, \ldots, \ell_T)$ 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

[1] 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.

[2] $|r|$ 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.

$H(\theta,\hat{\theta}) \leq H(\hat{\theta},\hat{\theta})$ 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 $\{f(\cdot|\theta_k)\}_{k=1,\ldots K}$ with mixture weights $\{\omega_k\}_{k=1,\ldots,K}$, 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:

$$c_{kt} = \frac{\hat{\omega}_k \mathcal{N}(x_t|\hat{m}_k,\hat{r}_k)}{\sum_{l=1}^{K}\hat{\omega}_l \mathcal{N}(x_t|\hat{m}_l,\hat{r}_l)} \quad (11)$$

$$c_k = \sum_{t=1}^{T} c_{kt} \quad (12)$$

$$\overline{x}_k = \sum_{t=1}^{T} c_{kt}x_t/c_k \quad (13)$$

$$S_k = \sum_{t=1}^{T} c_{kt}(x_t - \overline{x}_k)(x_t \cdot \overline{x}_k)^t. \quad (14)$$

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

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

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:

$$\nu'_k = \nu_k + c_k \quad (16)$$

$$\tau'_k = \tau_k + c_k \quad (17)$$

$$\alpha'_k = \alpha_k + c_k \quad (18)$$

$$\mu'_k = \frac{\tau_k\mu_k + c_k\overline{x}_k}{\tau_k + c_k} \quad (19)$$

$$u'_k = u_k + S_k + \frac{\tau_k c_k}{\tau_k + c_k}(\mu_k - \overline{x}_k)(\mu_k - \overline{x}_k)^t. \quad (20)$$

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 $\tilde{r}_k = (\alpha'_k - p)u'^{-1}_k$. Thus, the EM reestimation formulas are derived in (21)–(23), which appear at the bottom of this page.

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 MLE, 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 nondegenerate, i.e., $\hat{\omega}_k > 0$, then $c_{k1}, c_{k2}, \ldots, c_{kT}$ is a sequence of T i.i.d. random variables with a nondegenerate distribution and $\lim\sup_{T\to\infty}\sum_{t=1}^{T} c_{kt} = \infty$ with probability one [25]. It follows that $\tilde{w}_k$ converges to $\Sigma_{t=1}^{T} c_{kt}/T$ 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 $T \to \infty$.

IV. 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 $\theta_i = \{w_{ik}, m_{ik}, r_{ik}\}_{k=1,\ldots,K}$ for each state i.

For a sample $\mathbf{x} = (x_1,\ldots,x_T)$, the complete data is $\mathbf{y} = (\mathbf{x},\mathbf{s},\ell)$, where $\mathbf{s} = (s_0,\ldots,s_T)$ is the unobserved state sequence, and $\ell = (\ell_1,\ldots,\ell_T)$ 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 $\mathbf{x},\mathbf{s}$, and ℓ is defined

$$\tilde{\omega}_k = \frac{(\nu_k - 1) + \sum_{t=1}^{T} c_{kt}}{\sum_{k=1}^{K}(\nu_k - 1) + \sum_{k=1}^{K}\sum_{t=1}^{T} c_{kt}} \quad (21)$$

$$\tilde{m}_k = \frac{\tau_k\mu_k + \sum_{t=1}^{T} c_{kt}x_t}{\tau_k + \sum_{t=k}^{T} c_{kt}} \quad (22)$$

$$\tilde{r}_k^{-1} = \frac{u_k + \sum_{t=1}^{T} c_{kt}(x_t - \tilde{m}_k)(x_t - \tilde{m}_k)^t + \tau_k(\mu_k - \tilde{m}_k)(\mu_k - \tilde{m}_k)^t}{(\alpha_k - p) + \sum_{t=1}^{T} c_{kt}}. \quad (23)$$

as[3]

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

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 kth normal p.d.f. associated with state i. It follows that the likelihood of $\mathbf{x}$ has the form

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

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 $\mathbf{A}$ 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 (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 $\mathbf{A}$. This 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 $\{\pi_1, \ldots, \pi_N\}$, and $\{a_{i1}, \ldots, a_{iN}\}_{i=1,\ldots N}$. The prior density for all the HMM parameters thus satisfies the relation

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

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 (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 that were developed for ML estimation.

A. Forward-Backward MAP Estimate

From (24), it is straightforward to show that the auxiliary function of the EM algorithm applied to ML estimation of $\lambda, Q(\lambda, \hat{\lambda}) = E[\log h(y|\lambda)|\mathbf{x}, \hat{\lambda}]$ 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:

$$Q_\pi(\pi, \hat{\lambda}) = \sum_{i=1}^{N} \gamma_{i0} \log \pi_i \quad (27)$$

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

$$Q_A(\mathbf{A}, \hat{\lambda}) = \sum_{t=1}^{T} \sum_{i=1}^{T} \sum_{j=1}^{T} \Pr(s_{t-1} = i, s_t = j \mid \mathbf{x}, \hat{\lambda}) \log a_{ij} \quad (28)$$

$$= \sum_{i=1}^{N} Q_{a_i}(a_i, \hat{\lambda}) \quad (29)$$

$$Q_\theta(\theta, \hat{\lambda}) = \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{k=1}^{K} \Pr(s_t = i, \ell_t = k|\mathbf{x}, \hat{\lambda}) \log \omega_{ik} f(x_t|\theta_{ik}) \quad (30)$$

$$= \sum_{i=1}^{N} Q_{\theta_i}(\theta_i|\hat{\lambda}) \quad (31)$$

with

$$Q_{a_i}(a_i, \hat{\lambda}) = \sum_{t=1}^{T} \sum_{j=1}^{N} \xi_{ijt} \log a_{ij} \quad (32)$$

$$Q_{\theta_i}(\theta_i, \hat{\lambda}) = \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_{it} \frac{\hat{\omega}_{ik} f(x_t|\hat{\theta}_{ik})}{\sum_{l=1}^{K} \hat{\omega}_{il} f(x_t|\hat{\theta}_{il})} \log \omega_{ik} f(x_t|\theta_{ik}) \quad (33)$$

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 $\mathbf{x}$, 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 $\mathbf{x}$. 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: $\{\pi_1, \ldots, \pi_N\}, \{a_{i1}, \ldots, a_{iN}\}_{i=1,\ldots,N}$ and $\{\theta_i\}_{i=1,\ldots,N}$. 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 (11) is replaced by c_{ikt}, defined as

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

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 reestimation formulas for π and $\mathbf{A}$ 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 shown in (35)–(39), which appear at the bottom of the next page.

For multiple independent observation sequences $\{\mathbf{x}_v\}_{v=1,\ldots V}$, with $\mathbf{x}_v = \left(x_1^{(v)}, \ldots, x_{T_v}^{(v)}\right)$, we must maximize $G(\lambda) \prod_{v=1}^{V} f(\mathbf{x}_v|\lambda)$, where $f(\cdot|\lambda)$ is defined by (25). The EM auxiliary funciton 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 (24). It follows that the reestimation formulas for $\mathbf{A}$ and θ still hold if the summations over $t(\sum_{t=1}^{T})$ are replaced by summations over v and $t(\Sigma_{v=1}^{V})\Sigma_{t=1}^{T_v}$. 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

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

Just like for the mixture parameter case, it can be shown that as $V \rightarrow \infty$, 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 that 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).

B. 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 $\mathbf{s}$ $G(\lambda, \mathbf{s}|\mathbf{x})$, is maximized. The estimation procedure becomes

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

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 $\mathbf{s}$ and λ gives a sequence of estimates with nondecreasing 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

$$\mathbf{s}^{(m)} = \underset{\mathbf{s}}{\operatorname{argmax}}\, f\left(\mathbf{x}, \mathbf{s}|\lambda^{(m)}\right) \tag{42}$$

$$\lambda^{(m+1)} = \underset{\lambda}{\operatorname{argmax}}\, f\left(\mathbf{x}, \mathbf{s}^{(m)}|\lambda\right)G(\lambda). \tag{43}$$

The most likely state sequence $s^{(m)}$ 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 $f(\mathbf{x}, \mathbf{s}^{(m)}|\lambda^{(m+1)})G(\lambda^{(m+1)}) \geq f(\mathbf{x}, \mathbf{s}^{(m)}|\lambda^{(m)})G(\lambda^{(m)})$. 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 (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.

V. Prior Density Estimation

In the previous section, 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 stochatic 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 with 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 $\mathbf{X}$ is composed of multiple independent sequences associated with different

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

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

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

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

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

unknown values of the HMM parameters. Let $(\mathbf{X}, \Lambda) = [(\mathbf{x}_1, \lambda_1), (\mathbf{x}_2, \lambda_2), \ldots, (\mathbf{x}_Q, \lambda_Q)]$ 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 $f(\mathbf{X}|\varphi)$, which is 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, MLE 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 $\mathbf{X}$ 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.

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

$$\varphi^{(m+1)} = \underset{\varphi}{\operatorname{argmax}}\, G(\Lambda^{(m)}|\varphi). \tag{46}$$

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 MLE 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 (46) poses two problems. First, due to the Wishart and Dirichlet components, MLE 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 overparameterization 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 MLE's. For the overparameterization 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:

$$\nu_{ik} = \tau_{ik} + 1 \tag{47}$$

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

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 $\{\lambda_1, \lambda_2, \ldots, \lambda_Q\}$. 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 $\hat{\lambda} = \operatorname{argmax}_{\lambda} G(\lambda|\hat{\varphi})$ (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.

VI. 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] and 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 traning 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 (one to three utterances of each word), the MAP estimates gave better results than the MLE's.

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]–[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) and (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 V or can be arbitrarily fixed. For model adaptation, τ_{ik} can be regarded as a weight associated with the kth Gaussian of state i as shown in (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. Although several solutions have been investigated to smooth discrete HMM's, such as model interpolation, co-occurrence smoothing, and fuzzy VQ, only variance smoothing has been proposed for continuous density HMM's. In [10] and [11], we have shown that MAP estimation can be used to solve this problem for CDHMM's by tying the parameters of the prior density. Performance improvements 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 a mixture [11]. In experiments using the DARPA Naval Resources Management (RM) [24] and the *TI* connected digit corpora, MAP estimation always outperformed MLE 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]–[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 downsampled to 8 kHz to simulate the telephone bandwidth.

In Table I, 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 (seven males and five females). Results are reported using 40, 100, and 600 utterances (or equivalently about two, five, and 30 min of speech material) of the speaker-specific data (taken from RM SD data) for training and adaptation. The MLE (SD) and MAP (SI) word error rates using the standard RM word pair grammer are given in the two first rows of the table. The MLE (SD) word error rate for 2 min of training data is 31.5%. The SI word error rate (0 min of adaptation data) is 13.9%, which is somewhat comparable to the MLE result with 5 min of speaker-specific training data. Although 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 asymtotically similar as demonstrated in (35)–(39) with $T \rightarrow \infty$. Compared with the SI results, the word error reduction is 37% with 2 min 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.

TABLE I
SUMMARY OF SD, SA (SI), AND SA (M/F) RESULTS ON FEB91-SD TEST (Results are given as word error rate (%).)

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

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 I. The word error rate without speaker adaptation is 11.5%. The error rate is reduced to 7.5% with 2 min and 6.0% with 5 min of adaptation data. Comparing the last two 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 min of training data is 35% compared with the sex-dependent model results and 46% compared with the SI model results.

More details on experimental results using MAP estimation for parameter smoothing and model adaptation can be found in [10]–[12], [22] including application to speaker clustering and corrective training. MAP estimation has also has been applied to task adaptation [22]. In this case, task-independent SI models, which was trained from 10 000 utterances of 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.

VII. 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 and N-gram stochastic language models.

REFERENCES

[1] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occuring 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," in *Proc. ICASSP-83*, 1983, pp. 761–764.

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

[5] M. DeGroot, *Optimal Statistical Decisions*. New York: 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*, vol. 39, pp. 1–38, 1977.

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

[8] M. Ferretti and S. Scarci, "Large-vocabulary speech recognition with speaker-adapted codebook and HMM parameters," in *Proc. EuroSpeech-89*, 1989, pp. 154–156.

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

[10] J. -L. Gauvain and C. -H Lee, "Bayesian learning of gaussian mixture densities for hidden Markov models," in*Proc. DARPA Speech Natural Language Workshop* (Pacific Grove), Feb. 1991.

[11] ——, "MAP estimation of continuous density HMM: Theory and applications," *Proc. DARPA Speech Natural language Workshop* Feb. 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 Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, Mar. 1992.

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

[15] B. -H. Juang, "Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains," *AT&T Techn. J.*, vol. 64, no. 6, July-Aug. 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. 38, no. 9, Sept. 1990.

[17] B. O. Koopman, "On distributions admitting a sufficient statistic," *Trans. Amer. Math. Soc.*, vol. 39, pp. 399–409, 1936.

[18] L. F. Lamel and J. -L. Gauvain, "Cross-lingual experiments with phone recognition," in *Proc. IEEE ICASSP-93* (Minneapolis, MN), Apr. 1993, pp. 507-510, vol. 2.

[19] 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.

[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. 39, no. 4, pp. 806–814, Apr. 1991.

[21] C. -H. Lee, E. Giachin, L. R. Rabiner, R. Pieraccini, and A. E. Rosenberg, "Improved acoustic modeling for large vocabulary continous speech recognition," *Comput. Speech Language*, vol. 6, pp. 103–127, 1992.

[22] C. -H. Lee and J. -L. Gauvain, "Speaker adaptation based on map estimation of hmm parameters," in *Proc. IEEE ICASSP-93* (Minneapolis, MN), Apr. 1993, pp. 558-561, vol. 2..

[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, Sept. 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), Apr. 1988, pp. 651–654.

[25] Y. V. Prohorov and Y. A. Rozanov, *Probability Theory*. New York: 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.

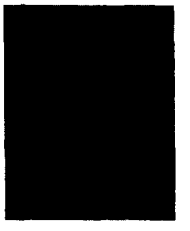
[27] C. R. Rao, *Linear Statistical Inference and Its Applications*. New York: Wiley, 1973, 2nd ed.

[28] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and em algorithm," *SIAM Rev.*, vol. 26, no. 2, pp. 195–239, Apr. 1984.

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

[30] R. Stern and M. Lasry, "Dynameic speaker adaptation for feature-based isolated word recognition," *IEEE Trans. Acoust. Speech Signal Processing*, 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," in *Proc. ICASSP-83* (Boston), 1983, pp. 1053–1056.



Jean-Luc Gauvain received the Master of Science degree in telecommunications in 1978 from the University of Paris XIII and the Third Cycle Docturate degree in electronics in 1982 from the University of Paris XI.

From October 1982 to September 1983, he was with the Vecsys Company, where he was involved in the development of the real-time CSR system Mozart, which was commercialized in 1983. Since October 1983, he has been a permanent CNRS researcher at LIMSI and is currently responsible for speech recognition research. His primary research centered on developing algorithms for speech processing, continuous speech recognition, and speaker verification. He has actively participated in the development of real-time prototypes for speech recognizers and speaker verification systems. He was also involved in the design and realization of the BREF corpus, which is a large French read-speech corpus of newspaper text. From June 1990 to November 1991, he was a visiting researcher at the Speech Research Department at AT&T Bell Laboratories. His current research interests are large vocabulary speech recognition and speaker and language identification.

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

In 1981, he joined Verbex Corp., Bedford, MA, and was involved in research work on connected word recognition. In 1984, he became affiliated with Digital Sound Corp., Santa Barbara, CA, where he was 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 Labs, Murray Hill, NJ. His current research interests include speech signal modeling, speech recognition, speaker recognition, and signal processing.

Since 1991, Dr. Lee has been an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING and is now an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He is a member of the DARPA Spoken Language Processing Coordination Committee.

High-Performance Connected Digit Recognition Using Maximum Mutual Information Estimation

Yves Normandin, Régis Cardin, and Renato De Mori, *Senior Member, IEEE*

***Abstract*— Hidden Markov Models (HMM's) are one of the most powerful speech recognition tools available today. Even so, the inadequacies of HMM's as a "correct" modeling framework for speech are well known. In this context, it is argued in this paper that the maximum mutual information estimation (MMIE) formulation for training is more appropriate than maximum likelihood estimation (MLE) for reducing the error rate.**

***Corrective MMIE training* is introduced. It is a very efficient new training algorithm which uses a modified version of a discrete reestimation formula recently proposed by Gopalakrishnan *et al.* Reestimation formulas are proposed for the case of diagonal Gaussian densities and their convergence properties are experimentally demonstrated. A description of how these formulas are integrated into our training algorithm is given. Using the MMIE framework for training, it is shown how weighting the contribution of different parameter sets in the computation of output probabilities introduces substantial recognition improvements.**

Using the TIDIGITS connected digit corpus, a large number of experiments are performed with the ideas, techniques, and algorithms presented in this paper. These experiments show that MMIE systematically provides substantial error rate reductions with respect to MLE alone and that, thanks to the new training techniques, these results can be obtained at an acceptable computational cost. The best results obtained in our experiments were 0.29% word error rate and 0.89% string error rate on the adult portion of the corpus.

I. Introduction

In automatic speech recognition (ASR) systems based on Hidden Markov Models (HMM's), the purpose of training is to find the HMM parameter set Θ which will result in the speech decoder with the lowest possible recognition error rate. The set Θ includes all transition probabilities and output distribution parameters in all HMM's used for a given task. Training is done by maximizing some objective function $R(\Theta)$. There are two important and difficult problems to consider. The first one is to determine a meaningful objective function. This function should be such that, whenever $R(\hat{\Theta}) > R(\Theta)$, then $\hat{\Theta}$ results in a better decoder than Θ. Once a function $R(\Theta)$ has been chosen, the second problem (the estimation problem) is to find the parameter set Θ that maximizes it.

By far the most common HMM parameter estimation technique is maximum likelihood estimation (MLE) [1]. Recently, a different type of estimation, called maximum mutual information estimation (MMIE) has been proposed [2]. There have been attempts at empirically justifying the use of MMIE with simple and well-controlled experiments. Some of them [3], [4] demonstrate that, for certain types of estimation problems, MMIE will converge to the optimal decoder even if incorrect modeling assumptions are made, while MLE will not. These experiments thus tend to show that MMIE is more robust than MLE when modeling assumptions are not correct. The fact that most of HMM's modeling assumptions about speech are incorrect could be an argument in favor of MMIE. In some cases, however, it is also possible that neither MMIE nor MLE will converge to the optimal decoder, but another type of estimator will.

It is not clear how these cases relate to speech recognition problems. In general, optimization algorithms will not converge to the global optimum and it is probably not possible to get an HMM-based optimal decoder for speech recognition. Thus the advantage of using MMIE for HMM-based ASR's should be assessed by experimentation. Many results reported in the literature [2], [4]–[6] tend to demonstrate MMIE's usefulness, but not conclusively.

We will show in this paper that, at least for the connected digit task on the TIDIGITS corpus [7], MMIE leads to significant recognition improvements with discrete and semicontinuous HMM's (SCHMM's). In a connected digit recognition experiment using one discrete model per digit, the string error rate was reduced from 1.92% to 1.48% by using MMIE after our standard MLE training. Further improvements (0.89% string error rate with two models per word) were obtained by using a new MMIE algorithm especially conceived for SCHMM's.

II. Relation Between MLE and MMIE

We assume that the result of a speaker pronouncing a word sequence (or message) $\mathbf{w}$ is an acoustic *observation sequence* $\mathbf{y} \equiv \underline{y}_1, \underline{y}_2, \cdots, \underline{y}_{L_{\mathbf{y}}}$. Typically, $\mathbf{y}$ is the result of a frame-based analysis performed on the speech signal produced by the speaker, where $\underline{y}_l$ is the parameter vector extracted from the lth frame. Let us assume that an HMM-type model can be built corresponding to any possible word sequence in the task, and let $\mathbf{m}_{\mathbf{w}}$ be the model corresponding to the word sequence $\mathbf{w}$. This model allows the computation of $P_\Theta(\mathbf{y}|\mathbf{m}_{\mathbf{w}})$, the probability that the model $\mathbf{m}_{\mathbf{w}}$ produced $\mathbf{y}$. Generally, $P_\Theta(\mathbf{y}|\mathbf{m}_{\mathbf{w}})$ is intended as an "estimate" of the probability $P(\mathbf{y}|\mathbf{w})$ that the pronunciation of the word sequence $\mathbf{w}$ resulted in $\mathbf{y}$. The reason for this is that the speech

Manuscript received October 10, 1991; revised April 15, 1993. The associated editor coordinating the review of this paper and approving it for publication was Dr. David Nahamoo.

The authors are with the Centre de recherche informatique de Montréal (CRIM), McGill College, Montréal, Québec, Canada, H3A 2N4.

IEEE Log Number 9215241.