MAP Estimation of Continuous Density HMM : Theory and Applications

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

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

ABSTRACT

We discuss *maximum a posteriori* estimation of continuous density hidden Markov models (CDHMM). The classical MLE reestimation algorithms, namely the forward-backward algorithm and the segmental k-means algorithm, are expanded and reestimation formulas are given for HMM with Gaussian mixture observation densities. Because of its adaptive nature, Bayesian learning serves as a unified approach for the following four speech recognition applications, namely parameter smoothing, speaker adaptation, speaker group modeling and corrective training. New experimental results on all four applications are provided to show the effectiveness of the MAP estimation approach.

INTRODUCTION

Estimation of hidden Markov model (HMM) is usually obtained by the method of maximum likelihood (ML) [1, 10, 6] assuming that the size of the training data is large enough to provide robust estimates. This paper investigates maximum a posteriori (MAP) estimate of continuous density hidden Markov models (CDHMM). The MAP estimate can be seen as a Bayes estimate of the vector parameter when the loss function is not specified [2]. This estimation technique provides a way of incorporatimg prior information in the training process, which is particularly useful to deal with problems posed by sparse training data for which the ML approach gives inaccurate estimates. This approach can be applied to two classes of estimation problems, namely, parameter smoothing and model adaptation, both related to the problem of sparse training data.

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

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, i.e.

$$\theta_{\text{MAP}} = \arg\max_{\theta} f(\mathbf{x}|\theta)\, g(\theta) \qquad (1)$$

If θ is assumed to be fixed but unknown, then there is no knowledge about θ, which is equivalent to assuming a non-informative improper prior, i.e. $g(\theta) =$constant. Equation (1) then reduces to the familiar ML formulation.

Given the MAP formulation two problems remain: the choice of the prior distribution family and the evaluation of the maximum a posteriori. These two problems are closely related, since the appropriate choice of the prior distribution can greatly simplify the MAP estimation. Like for 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 $t(\mathbf{x})$. 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 $f(\cdot|\theta)$, i.e. $\forall \mathbf{x}\ t(\mathbf{x}) \in \phi$ [4, 2]. The MAP estimation is then reduced to the evaluation of the mode of $k(\theta|\varphi') = k(\theta|\varphi)k(\theta|t(\mathbf{x}))$, a problem almost identical to the ML estimation problem. However, among the families of interest, only exponential families have a sufficient statistic of fixed dimension [7]. When there is no sufficient statistic of 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 density and hidden Markov model, the lack of a sufficient statistic of fixed dimension is due to the underlying hidden process, i.e. a multinomial model for the mixture and a Markov chain for an HMM. In these cases ML estimates are usually obtained by using the expectation-maximization (EM) algorithm [3, 1, 13]. This algorithm exploits the fact that the complete-data likelihood can be 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. [3], the EM algorithm can also be applied to MAP estimation. In the next two sections the formulations of this algorithm for MAP estimation of Gaussian mixture and CDHMM with Gaussian mixture observation densities are derived.

MAP ESTIMATES FOR GAUSSIAN MIXTURE

Suppose that $\mathbf{x} = (x_1, ..., x_n)$ is a sample of n i.i.d. observations drawn from a mixture of K p-dimensional multivariate normal densities. The joint p.d.f. is specified by $f(\mathbf{x}|\theta) = \prod_{t=1}^{n} \sum_{k=1}^{K} \omega_k \mathcal{N}(x_t|m_k, r_k)$ where $\theta = (\omega_1, ..., \omega_K, m_1, ..., m_K, r_1, ..., r_K)$ is the parameter vector and ω_k denotes the mixture gain for the k-th mixture component with the constraint $\sum_{k=1}^{K} \omega_k = 1$. $\mathcal{N}(x|m_k, r_k)$ is the k-th normal density function where m_k is the p-dimensional mean vector and r_k is the $p \times p$ precision matrix. As stated in the introduction, for the parameter vector θ no joint conjugate prior density exists. 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 product of a multinomial density (for the sizes of the component populations) and normal densities (for the component densities). A practical candidate to model the

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

prior knowledge about the mixture gain parameter vector is therefore a Dirichlet density which is the conjugate prior density for the multinomial distribution

$$g(\omega_1, ..., \omega_K) \propto \prod_{k=1}^{K} \omega_k^{\nu_k - 1} \tag{2}$$

where $\nu_k > 0$. For the vector parameter (m_k, r_k) of the individual Gaussian mixture component, the joint conjugate prior density is a normal-Wishart density [2] of the form

$$g(m_k, r_k) \propto |r_k|^{(\alpha_k - p)/2} \exp[-\tfrac{1}{2}\mathrm{tr}(u_k r_k)] \times \exp[-\frac{\tau_k}{2}(m_k - \mu_k)^T r_k (m_k - \mu_k)] \tag{3}$$

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 mixture components and the mixture weights, the joint prior density $g(\theta)$ is taken to be a product of the prior p.d.f.'s defined in equations (2) and (3), i.e. $g(\theta) = g(\omega_1, ..., \omega_K) \prod_{k=1}^{K} g(m_k, r_k)$. As will be shown later, 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 is in the conjuguate family of the *complete-data* density.

The EM algorithm is an iterative procedure for approximating maximum-likelihood estimates in an incomplete-data context such as mixture density and hidden Markov model estimation problems [1, 3, 13]. This procedure consists of maximizing at each iteration the auxilliary function $Q(\theta, \hat{\theta})$ defined as the expectation of the *complete-data* log-likelihood $\log h(\mathbf{y}|\theta)$ given the incomplete data $\mathbf{x} = (x_1, ..., x_n)$ and the current fit $\hat{\theta}$, i.e. $Q(\theta, \hat{\theta}) = E[\log h(\mathbf{y}|\theta)|\mathbf{x}, \hat{\theta}]$. For a mixture density, the complete-data likelihood is the joint likelihood of $\mathbf{x}$ and $\ell = (\ell_1, ..., \ell_n)$ the unobserved labels referring to the mixture components, i.e. $\mathbf{y} = (\mathbf{x}, \ell)$.

The EM procedure derives from the fact 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 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 auxilliary function $R(\theta, \hat{\theta}) = Q(\theta, \hat{\theta}) + \log g(\theta)$ at each iteration instead of $Q(\theta, \hat{\theta})$ [3].

For a mixture of K densities $\{f(\cdot|\theta_k)\}_{k=1,...,K}$ with mixture weights $\{\omega_k\}_{k=1,...,K}$, the auxilliary function Q takes the following form [13]:

$$Q(\theta, \hat{\theta}) = \sum_{t=1}^{n} \sum_{k=1}^{K} \frac{\hat{\omega}_k f(x_t|\hat{\theta}_k)}{f(x_t|\hat{\theta})} \log \omega_k f(x_t|\theta_k) \tag{4}$$

Let $\Psi(\theta, \hat{\theta}) = \exp R(\theta, \hat{\theta})$ be the function to be maximized and define the following notations $c_{kt} = \dfrac{\hat{\omega}_k f(x_t|\hat{\theta}_k)}{f(x_t|\hat{\theta})}$, $c_k = \sum_{t=1}^{n} c_{kt}$, $\bar{x}_k = \sum_{t=1}^{n} c_{kt} x_t / c_k$ and $S_k = \sum_{t=1}^{n} c_{kt}(x_t - \bar{x}_k)(x_t - \bar{x}_k)^T$. It follows from the definition of $f(\mathbf{x}|\theta)$ and equation (4) that

$$\Psi(\theta, \hat{\theta}) \propto g(\theta) \prod_{k=1}^{K} \omega_k^{c_k} |r_k|^{c_k/2} \times \exp[-\frac{c_k}{2}(m_k - \bar{x}_k)^T r_k (m_k - \bar{x}_k) - \tfrac{1}{2}\mathrm{tr}(S_k r_k)] \tag{5}$$

From (2), (3) and (5) it can easily be verified that $\Psi(\cdot, \hat{\theta})$ belongs to the same family as g, and has parameters $\{\nu_k', \tau_k', \mu_k', \alpha_k', u_k'\}_{k=1,...,K}$ satisfying the following conditions:

$$\nu_k' = \nu_k + c_k \tag{6}$$

$$\tau_k' = \tau_k + c_k \tag{7}$$

$$\alpha_k' = \alpha_k + c_k \tag{8}$$

$$\mu_k' = \frac{\tau_k \mu_k + c_k \bar{x}_k}{\tau_k + c_k} \tag{9}$$

$$u_k' = u_k + S_k + \frac{\tau_k c_k}{\tau_k + n_k}(\mu_k - \bar{x}_k)(\mu_k - \bar{x}_k)^T \tag{10}$$

The considered family of distributions is therefore a conjugate family for the complete-data density.

The mode of $\Psi(\cdot, \hat{\theta})$, denoted (ω_k', m_k', r_k'), may be obtained from the modes of the Dirichlet and normal-Wishart densities: $\omega_k' = (\nu_k' - 1)/\sum_{k=1}^{K}(\nu_k' - 1)$, $m_k' = \mu_k'$, and $r_k' = (\alpha_k' - p)u_k'^{-1}$. Thus, the EM iteration is as follows:

$$\omega_k' = \frac{\nu_k - 1 + \sum_{t=1}^{n} c_{kt}}{n - K + \sum_{k=1}^{K} \nu_k} \tag{11}$$

$$m_k' = \frac{\tau_k \mu_k + \sum_{t=1}^{n} c_{kt} x_t}{\tau_k + \sum_{t=1}^{n} c_{kt}} \tag{12}$$

$$r_k'^{-1} = \frac{u_k + \tau_k(\mu_k - m_k')(\mu_k - m_k')^T}{\alpha_k - p + \sum_{t=1}^{n} c_{kt}} + \frac{\sum_{t=1}^{n} c_{kt}(x_t - m_k')(x_t - m_k')^T}{\alpha_k - p + \sum_{t=1}^{n} c_{kt}} \tag{13}$$

If it is assumed $\hat{\omega}_k > 0$, then $c_{k1}, c_{k2}, ..., c_{kn}$ is a sequence of n i.i.d. random variables with a non-degenerate distribution and $\limsup_{n\to\infty} \sum_{t=1}^{n} c_{kt} = \infty$ with probability one. It follows that w_k' converges to $\sum_{t=1}^{n} c_{kt}/n$ with probability one when $n \to \infty$. Applying the same reasoning to m_k' and 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 are identical, the EM algorithm will provide identical estimates with probability one when $n \to \infty$.

MAP ESTIMATES FOR CDHMM

The results obtained for a mixture of normal densities can be extended to the case of HMM with Gaussian mixture state observation densities, assuming that the observation p.d.f.'s of all the states have the same number of mixture components. We 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 $\mathbf{x} = (x_1, ..., x_n)$, the complete data is $\mathbf{y} = (\mathbf{x}, \mathbf{s}, \ell)$ where $\mathbf{s} = (s_0, ..., s_n)$ is the unobserved state sequence, and $\ell = (\ell_1, ..., \ell_n)$ are the unobserved mixture component labels, $s_i \in [1, N]$ and $l_i \in [1, K]$. The joint p.d.f. $h(\cdot|\lambda)$ of $\mathbf{x}$, $\mathbf{s}$, and ℓ is defined as [1]

$$h(\mathbf{x}, \mathbf{s}, \ell|\lambda) = \pi_{s_0} \prod_{t=1}^{n} a_{s_{t-1}s_t} \omega_{s_t \ell_t} f(x_t|\theta_{s_t \ell_t}) \tag{14}$$

where π_i is the initial probabilty 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 to state i. It follows that the likelihood of $\mathbf{x}$ has the form

$$f(\mathbf{x}|\lambda) = \sum_{\mathbf{S}} \pi_{s_0} \prod_{t=1}^{n} a_{s_{t-1}s_t} f(x_t|\theta_{s_t}) \tag{15}$$

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.

In the general case where MAP estimation is to be applied not only to the observation density parameters but also to the initial and transition probabilities, a Dirichlet density can also be used for the initial probability vector π and for each row of the transition probability matrix $\mathbf{A}$. This choice directly follows the results of 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 parameters $\{n, \pi_1, ..., \pi_N\}$ and $\{n, a_{i1}, ..., a_{iN}\}_{i=1,...,N}$. The prior density for all the HMM parameters is thus

$$G(\lambda) \propto \prod_{i=1}^{N} \left[\pi_i^{\eta_i - 1} g(\theta_i) \prod_{j=1}^{N} a_{ij}^{\eta_{ij} - 1} \right] \tag{16}$$

In the following subsections we examine two ways of approximating λ_{MAP} by local maximization of $f(\mathbf{x}|\lambda)G(\lambda)$ and $f(\mathbf{x}, \mathbf{s}|\lambda)G(\lambda)$. These two solutions are the MAP versions of the Baum-Welch algorithm [1] and of the segmental k-means algorithm [12], algorithms which were developed for ML estimation.

Forward-Backward MAP Estimate

From (14) it is straightforward to show that the auxilliary function of the EM algorithm applied to MLE of λ, $Q(\lambda, \hat{\lambda}) = E[\log h(\mathbf{y}|\lambda)|\mathbf{x}, \hat{\lambda}]$, can be decomposed into a sum of three auxilliary functions: $Q_\pi(\pi, \hat{\lambda})$, $Q_A(\mathbf{A}, \hat{\lambda})$ and $Q_\theta(\theta, \hat{\lambda})$ [6]. These functions which can be independently maximized take the following forms:

$$Q_\pi(\pi, \hat{\lambda}) = \sum_{i=1}^{N} \gamma_{i0} \log \pi_i \tag{17}$$

$$Q_A(\mathbf{A}, \hat{\lambda}) = \sum_{i=1}^{N} \sum_{t=1}^{n} \sum_{j=1}^{N} \xi_{ijt} \log a_{ij} \tag{18}$$

$$Q_\theta(\theta, \hat{\lambda}) = \sum_{i=1}^{N} Q_{\theta_i}(\theta_i|\hat{\lambda}) \tag{19}$$

with

$$Q_{\theta_i}(\theta_i, \hat{\lambda}) = \sum_{t=1}^{n} \sum_{k=1}^{K} \gamma_{it} \frac{\hat{\omega}_{ik} f(x_t|\hat{\theta}_{ik})}{f(x_t|\hat{\theta}_i)} \log \omega_{ik} f(x_t|\theta_{ik}) \tag{20}$$

where $\xi_{ijt} = \Pr(s_{t-1} = i, s_t = j|\mathbf{x}, \hat{\lambda})$ and $\gamma_{it} = \Pr(s_t = i|\mathbf{x}, \hat{\lambda})$ can be computed at each EM iteration by using the Forward-Backward algorithm [1]. As for the mixture Gaussian case discussed in the previous section, to estimate the mode of the posterior density the auxilliary function $R(\lambda, \hat{\lambda}) = Q(\lambda, \hat{\lambda}) + \log G(\lambda)$ must be maximized. The form chosen for $G(\lambda)$ in (16) 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 auxilliary 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 auxilliary function associated with the indexed parameter set.

We can recognize in (20) the same form as seen for $Q(\theta|\hat{\theta})$ in (4) for the mixture Gaussian case. It follows that if the c_{kt} are replaced by the c_{ikt} defined as

$$c_{ikt} = \gamma_{it} \frac{\hat{\omega}_{ik} \mathcal{N}(x_t|\hat{m}_{ik}, \hat{r}_{ik})}{f(x_t|\hat{\theta}_i)} \tag{21}$$

then the reestimation formulas (11-13) can be used to maximize $R_{\theta_i}(\theta_i, \hat{\lambda})$. It is straightforward to find the reestimations formulas for π and $\mathbf{A}$ by applying the same derivations used for the mixture weights:

$$\pi_i' = \frac{\eta_i - 1 + \gamma_{i0}}{\sum_{j=1}^{N} \eta_j - N + \sum_{j=1}^{N} \gamma_{j0}} \tag{22}$$

$$a_{ij}' = \frac{\eta_{ij} - 1 + \sum_{t=1}^{n} \xi_{ijt}}{\sum_{j=1}^{N} \eta_{ij} - N + \sum_{j=1}^{N} \sum_{t=1}^{n} \xi_{ijt}} \tag{23}$$

For multiple independent observation sequences $\{\mathbf{x}_q\}_{q=1,...,Q}$, with $\mathbf{x}_q = (x_1^{(q)}, ..., x_{n_q}^{(q)})$, we maximize $G(\lambda) \prod_{q=1}^{Q} f(\mathbf{x}_q|\lambda)$, where $f(\cdot|\lambda)$ is defined by (15). The EM auxilliary function is then $R(\lambda, \hat{\lambda}) = \log G(\lambda) + \sum_{q=1}^{Q} E[\log h(\mathbf{y}_q|\lambda)|\mathbf{x}_q, \hat{\lambda}]$, where $h(\cdot|\lambda)$ is defined by equation (14). It follows that the reestimation formulas for $\mathbf{A}$ and θ still hold if the summations over t are replaced by summations over q and t. The values $\xi_{ijt}^{(q)}$ and $\gamma_{it}^{(q)}$ are then obtained by applying the forward-backward algorithm for each observation sequence. The reestimation formula for the initial probabilities becomes

$$\pi_i' = \frac{\eta_i - 1 + \sum_{q=1}^{Q} \gamma_{i0}^{(q)}}{\sum_{j=1}^{N} \eta_j - N + \sum_{q=1}^{Q} \sum_{j=1}^{N} \gamma_{j0}^{(q)}} \tag{24}$$

As for the mixture Gaussian case, it can be shown that as $Q \to \infty$, the MAP reestimation formulas approach the ML ones, exhibiting the asymptotic 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 essential to finding a solution close to a global maximum and to minimize the number of EM iterations needed to attain the local maximum. When using an informative prior, one 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 n equal to 0. When using a non-informative prior, i.e. for ML estimation, while for discrete HMMs it is possible to use uniform initial estimates, there is no trivial solution for the continuous density case.

Segmental MAP Estimate

By analogy with the segmental k-means algorithm [12], a different optimization criterion can be considered. Instead of maximizing $G(\lambda|\mathbf{x})$, the joint posterior density of λ and $\mathbf{s}$, $G(\lambda, \mathbf{s}|\mathbf{x})$, is maximized. The estimation procedure becomes

$$\tilde{\lambda} = \operatorname*{argmax}_{\lambda} \max_{\mathbf{S}} G(\lambda, \mathbf{s}|\mathbf{x}) \tag{25}$$

$$= \operatorname*{argmax}_{\lambda} \max_{\mathbf{S}} f(\mathbf{x}, \mathbf{s}|\lambda) G(\lambda) \tag{26}$$

and $\tilde{\lambda}$ is called the *segmental MAP estimate* of λ. As for the segmental k-means algorithm, 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 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

$$\mathbf{s}^{(m)} = \operatorname*{argmax}_{\mathbf{s}} f(\mathbf{x}, \mathbf{s}|\lambda^{(m)}) \quad (27)$$

$$\lambda^{(m+1)} = \operatorname*{argmax}_{\lambda} f(\mathbf{x}, \mathbf{s}^{(m)}|\lambda) G(\lambda) \quad (28)$$

The most likely state sequence $\mathbf{s}^{(m)}$ is decoded by the Viterbi algorithm. In fact, maximization over λ can be replaced by any *hill climbing* procedure which replaces $\lambda^{(m)}$ by $\lambda^{(m+1)}$ 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 auxilliary 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 (14). It is straightforward to show that the forward-backward reestimation equations 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.

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 (16). 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. Another solution is to adopt an empirical Bayesian approach [14] where the prior parameters are estimated directly from data. The estimation is then based on the marginal distribution of the data given the prior parameters.

Adopting the empirical Bayes approach, it is assumed that the sequence of observations, $\mathbf{X}$, is composed of multiple independent sequences associated with different unknown values of the HMM parameters. Letting $(\mathbf{X}, \Lambda) = [(\mathbf{x}_1, \lambda_1), (\mathbf{x}_2, \lambda_2), ...]$ 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 $f(\mathbf{X}|\varphi)$,

$$f(\mathbf{X}|\varphi) = \int_{\Lambda} f(\mathbf{X}|\Lambda) G(\Lambda|\varphi)\, d\Lambda \quad (29)$$

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 simplify this problem, we can choose a simpler optimization criterion by maximizing the joint p.d.f. $f(\mathbf{X}, \Lambda|\varphi)$ over Λ and φ instead of the marginal p.d.f. of $\mathbf{X}$ given φ. Starting with an initial estimate of φ, we obtain a hill climbing procedure by alternate maximization over Λ and φ, i.e.

$$\Lambda^{(m)} = \operatorname*{argmax}_{\Lambda} f(\mathbf{X}, \Lambda|\varphi^{(m)}) \quad (30)$$

$$\varphi^{(m+1)} = \operatorname*{argmax}_{\varphi} G(\Lambda^{(m)}|\varphi) \quad (31)$$

Such a procedure provides a sequence of estimates with non-decreasing values of $f(\mathbf{X}, \Lambda|\varphi^{(m)})$. The solution of (30) is the MAP estimate of Λ based on the current prior parameter $\varphi^{(m)}$. It can therefore be obtained by applying the forward-backward MAP reestimation formulas to each observation sequence $\mathbf{x}_q$. The solution of (31) is simply the maximum likelihood estimate of φ based on the current values of the HMM parameters.

Finding this estimate poses two problems. First, due to the Wishart and Dirichlet components, ML estimation for the density defined by (16) 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 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 can be verified that the following constraints hold

$$\nu_{ik} = \tau_{ik} \quad (32)$$

$$\alpha_{ik} = \tau_{ik} + p \quad (33)$$

Parameter tying can also be used to further reduce the size of the prior family.

We use this approach for approach for two types of applications: parameter smoothing and adaptation learning. For parameter "smoothing", the goal is to estimate $\{\lambda_1, \lambda_2, ...\}$. The previous algorithm offers a direct solution to "smooth" these different estimates by assuming a common prior density for all the models. For adaptative learning, we observe a new sequence of observations $\mathbf{x}_q$ associated with the unobserved vector parameter value λ_q. The MAP estimate of λ_q can be obtained by using for prior parameters a point estimate $\hat{\varphi}$ obtained with the previous algorithm. Such a training process can be seen as an adaptation of an *a priori* model $\hat{\lambda} = \operatorname{argmax}_{\lambda} G(\lambda|\hat{\varphi})$ (when no training data is available) to more specific conditions corresponding to the new observation sequence $\mathbf{x}_q$.

In the applications presented in this paper, the prior density parameters were estimated along with the estimation of the SI model parameters using the segmental k-means algorithm. Information about the variability to be modeled with the prior densities was associated with each frame of the SI training data. This information was simply represented by a class number which can be the speaker ID, the speaker sex, or the phonetic context. The HMM parameters for each class given the mixture component were then computed, and moment estimates were obtained for the tied prior parameters also subject to conditions (32-33) [5].

EXPERIMENTAL SETUP

The experiments presented in this paper used various sets of context-independent (CI) and context-dependent (CD) phone models. Each model is a left-to-right HMM with Gaussian mixture state observation densities. Diagonal covariance matrices are used and the transition probabilities are assumed fixed and known. As described in [8], a 38-dimensional feature vector composed of LPC-derived cepstrum coefficients, and first and second order time derivatives. Results are reported for the RM task with the standard word pair grammar and for the TI/NIST connected digits. Both corpora were down-sampled to telephone bandwidth.

MODEL SMOOTHING AND ADAPTATION

Last year we reported results for CD model smoothing, speaker adaptation, and sex-dependent modeling [5]. CD model smoothing was found to reduce the word error rate by 10%. Speaker adaptation

Training	0 min	2 min	5 min	30 min
SD	—-	31.5	12.1	3.5
SA (SI)	13.9	8.7	6.9	3.4
SA (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 (%).

was tested on the JUN90 data with 1 minute and 2 minutes of speaker-specific adaptation data. A 16% and 31% reduction in word error were obtained compared to the SI results [5]. On the FEB91 test, using Bayesian learning for CD model smoothing combined with sex-dependent modeling, a 21% word error reduction was obtained compared to the baseline results [5].

In order to compare speaker adaption to ML training of SD models, an experiment has been carried out on the FEB91-SD test material including data from 12 speakers (7m/5f), using a set of 47 CI phone models. Two, five and thirty minutes of the SD training data were used for training and adaptation. The SD, SA (SI) word error rates are given in the two first rows of Table 1.

The SD word error rate for 2 min of training data was 31.5%. The SI word error rate (0 minutes of adaptation data) was 13.9%, somewhat comparable to the SD results with 5 min of SD training data. The SA models are seen to perform better than SD models when relatively small amounts of data were used for training or adaptation. When all the available training data was used, the SA and SD results were comparable, consistent with the Bayesian formulation that the MAP estimate converges to the MLE. Relative to the SI results, the word error reduction was 37% with 2 min of adaptation data, an improvement similar to that observed on the JUN90 test data with CD models [5]. As in the previous experiment, a larger improvement was observed for the female speakers (51%) than for the male speakers (22%).

Speaker adaptation was also performed starting with sex-dependent models (third row of Table 1). The word error rate with no 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 2 rows of the table it can be seen that SA is more effective when sex-dependent seed models are used. The error reduction with 2 min of training data is 35% compared to the sex-dependent model results and 46% compared to the SI model results.

P.D.F. SMOOTHING

We have shown that Bayesian learning can be used for CD model smoothing [5]. This approach can be seen either as a way to add extra constraints to the model parameters so as to reduce the effect of insufficient training data, or it can be seen as an "interpolation" between two sets of parameter estimates: one corresponding to the desired model and the other to a smaller model which can be trained using MLE on the same data. Instead of defining a reduced parameter set by removing the context dependency, we can alternatively reduce the mixture size of the observation densities and use a single Gaussian per state in the smaller model. Cast in the Bayesian learning framework, this implies that the same marginal prior density is used for all the components of a given mixture. Variance clipping can also be viewed as a MAP estimation technique with a uniform prior density constrained by a maximum (positive) value for the precision parameters [9]. However, this does not have the appealing interpolation capability of the conjugate priors.

We experimented with this p.d.f. smoothing approach on the TI

	WACC	SACC (Strings Correct)
MLE	99.6	98.7 (8464)
MLE+VC	99.6	98.8 (8477)
MAP	99.7	99.1 (8502)

Table 2: TI test results for p.d.f. smoothing (213 inter-word CD-32 models)

	FEB89	OCT89	JUN90	FEB91
MLE	93.3	92.5	92.1	92.9
MLE+VC	95.0	95.0	94.8	95.9
MAP(SI)	95.0	95.5	95.0	96.2
MAP(M/F)	95.2	96.2	95.2	96.7

Table 3: RM test results for p.d.f. smoothing (2421 inter-word CD-16 models

digit and RM databases. A set of 213 CD phone models with 32 mixture components (213 CD-32) for the TI digits and a set of 2421 CD phone models with 16 mixture components (2421 CD-16) for RM were used for evaluation. Results are given for MLE training, MLE with variance clipping (MLE+VC), and MAP estimation with p.d.f. smoothing in Tables 2 and 3. In Table 2, word accuracy (WACC) and string accuracy (SACC) are given for the 8578 test digit strings of the TI digit corpora. Compared to the variance clipping scheme, the MAP estimate reduces the number of string errors by 25%. Using p.d.f. smoothing, the string accuracy of 99.1% is the best result reported on this task.

For the RM tests summarized in Table 3, a consistent improvement over the variance clipping scheme (MLE+VC) is observed when p.d.f. smoothing is applied. Combined with sex-dependent modeling, the MAP(M/F) scheme gives an average word accuracy of about 95.8%.

CORRECTIVE TRAINING

Bayesian learning provides a scheme for model adaptation which can also be used for corrective training. Corrective training maximizes the recognition rate on the training data hoping that that will also improve performance on the test data. One simple way to do corrective training is to use the training sentences which were incorrectly recognized as new data. In order to do so, the state segmentation step of the segmental MAP algorithm was modified to obtain not only the frame/state association for the *sentence model* states but also for the states corresponding to the model of all the possible sentences (*general model*). In the reestimation formulas, the values c_{ikt} for each state s_i are evaluated using (21), such that γ_{it} is equal to 1 in the sentence model and to -1 in the general model. While convergence is not guaranteed, in practice it was found that by using large values for $\tau_{ik} (\simeq 200)$, the number of training sentence errors decreased after each iteration until convergence. If we use the forward-backward MAP algorithm we obtain a corrective training algorithm for CDHMM's very similar to the recently proposed *corrective MMIE training* algorithm [11].

Corrective training was evaluated on both the TI/NIST SI connected digit and the RM tasks. Only the Gaussian mean vectors and the mixture weights were corrected. For the TI digits a set of 21 phonetic HMMs were trained on the 8565 digit strings. Results are given in Table 4 using 16 and 32 mixture components for the observation p.d.f.'s, with and without corrective training for both test and training data. The CT-16 results were obtained with 8 iter-

Training Conditions	Training		Test	
	string	word	string	word
MLE-16	1.6 (134)	0.5	2.0 (168)	0.7
CT-16	0.2 (18)	0.1	1.4 (122)	0.5
MLE-32	0.8 (67)	0.2	1.5 (126)	0.5
CT-32	0.3 (29)	0.1	1.3 (111)	0.4

Table 4: Corrective training results in string and word error rates (%) on the TI-digits for 21 CI models with 16 and 32 mixture components per state. String error counts are given in parenthesis.

Test Set	MLE-32	CT-32	ICT-32
TRAIN	7.7	1.8	3.1
FEB89	11.9	10.2	8.9
OCT89	11.5	9.8	8.9
JUN90	10.2	8.8	8.1
FEB91	11.4	10.3	10.2
FEB91-SD	13.9	11.3	11.0
Overall Test	11.8	10.1	9.4

Table 5: Corrective training results on the RM task (47 CI models with 32 mixture components per state)

ations of corrective training while the CT-32 results were based on only 3 iterations, where one full iteration of corrective training is implemented as one recognition run which produces a set of "new" training strings (i.e. errors and/or barely correct strings) followed by 10 iterations of Bayesian adaptation using the data of these strings. String error rates of 1.4% and 1.3% were obtained with 16 and 32 mixture components per state respectively, compared to 2.0% and 1.5% without corrective training. These represent string error reductions of 27% and 12%. We note that corrective training helps more with smaller models, as the ratio of adaptation data to the number of parameters is larger.

The corrective training procedure is also effective for continuous sentence recognition of the RM task. Table 5 gives results for the RM task, using 47 SI-CI models with 32 mixture components. The CT-32 corrective training assumes a fixed beam width. Since the number of string errors was small in the training set, the amount of data for corrective training was rather limited. To increase the amount, a smaller beam width was used to recognize the training data. It was observed that this *improved corrective training* (ICT-32) procedure not only reduced the error rate in training but also increased the separation between the correct string and the other competing strings. The number of training errors also increased as predicted. The regular and the improved corrective training gave an average word error rate reduction of 15% and 20% respectively on the test data.

SUMMARY

The theoretical framework for MAP estimation of multivariate Gaussian mixture density and HMM with mixture Gaussian state observation densities was presented. Two MAP training algorithms, the *forward-backward MAP estimation* and the *segmental MAP estimation*, were formulated. Bayesian learning serves as a unified approach for speaker adaptation, speaker group modeling, parameter smoothing and corrective training.

Tested on the RM task, encouraging results have been obtained for all four applications. For speaker adaptation, a 37% word error reduction over the SI results was obtained on the FEB91-SD test with 2 minutes of speaker-specific training data. It was also found that speaker adaptation is more effective when based on sex-dependent models than with an SI seed. Compared to speaker-dependent training, speaker adaptation achieved a better performance with the same amount of training/adaptation data. Corrective training applied to CI models reduced word errors by 15-20%. The best SI results on RM tests were obtained with p.d.f. smoothing and sex-dependent modeling, an average word accuracy of about 95.8% on four test sets.

Only corrective training and p.d.f. smoothing were applied to the TI/NIST connected digit task. It was found that corrective training is effective for improving CI models, reducing the number of string errors by up to 27%. Corrective training was found to be more effective for models having smaller numbers of parameters. This implies that we can reduce computational requierements by using corrective training on a smaller model and achieve performance comparable to that of a larger model. Using 213 CD models, p.d.f. smoothing provided a robust model that gave a 99.1% string accuracy on the test data, the best performance reported on this corpus.

REFERENCES

[1] 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.

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

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

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

[5] 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, Feb. 1991.

[6] 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.

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

[8] C.-H. Lee, E. Giachin, L. R. Rabiner, R. Pieraccini and A. E. Rosenberg, "Improved Acoustic Modeling for Continuous Speech Recognition", *Proc. DARPA Speech and Natural Language Workshop*, Hidden Valley, June 1990.

[9] 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. on ASSP*, April 1991.

[10] 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.

[11] Y. Normandin and D. Morgera, "An Improved MMIE Training Algorithm for Speaker-Independent Small Vocabulary, Continuous Speech Recognition", *Proc. ICASSP91*, pp. 537-540, May 1991.

[12] 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.

[13] 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.

[14] H. Robbins, "The Empirical Bayes Approach to Statistical Decision Problems," *Ann. Math. Statist.*, Vol. 35, pp. 1-20, 1964.