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

Qiang Huo[†‡] *and Chorkin Chan*[†]

[†]Department of Computer Science
The University of Hong Kong, Hong Kong

[‡]Department of Radio and Electronics
University of Science and Technology of China, P.R.C.

Abstract

In this paper a theoretical framework for Bayesian adaptive learning of discrete HMM and semi-continuous one with Gaussian mixture state observation densities is presented. Corresponding to the well-known Baum-Welch and segmental k-means algorithms respectively for HMM training, formulations of MAP (maximum *a posteriori*) and segmental MAP estimation of HMM parameters are developed. Furthermore, a computationally efficient method of the segmental quasi-Bayes estimation for semi-continuous HMM is also presented. The important issue of prior density estimation is discussed and a simplified method of moment estimate is given. The method proposed in this paper will be applicable to some problems in HMM training for speech recognition such as sequential or batch training, model adaptation, and parameter smoothing, etc.

Keywords: Bayesian learning; empirical Bayes method; Hidden Markov model; automatic speech recognition; speaker adaptation; parameter smoothing.

1 Introduction

The use of hidden Markov models (HMMs) for speech recognition has become increasingly popular in the past few years. The widespread popularity of the HMM framework can mainly be attributed to the existence of the efficient training procedures for HMM. Among these algorithms, the Baum-Welch [2, 3, 20, 15, 16] and segmental k-means [24, 17], are two most commonly used procedures for the estimation of HMM parameters. By assuming the HMM parameters to be fixed but unknown, these parameter estimators have been derived purely from the training observation sequences (sample information) plus some constraints that these parameters must obey without any prior information included. There may be many cases in which the prior information about the HMM parameters is available. Such information may, for example, come from subject matter considerations and/or previous experience. If indeed such information is available, the investigator may wish to use it in addition to the sample information in making inference about the HMM parameters. As is well known, the Bayesian inference approach provides a convenient method for combining sample and prior information. By assuming the HMM parameters to be random, this prior information is expressed in the form of a prior distribution, which is combined with a likelihood function via Bayes' theorem to form a posterior distribution on which inferences are based. Consequently, the flexibility in incorporating varying amount of prior information makes the Bayesian inference procedure successful in handling the problem of limited amount of relevant sampling data as well as applicable to certain problems of HMM training for speech recognition such as sequential or batch training, model adaptation, parameter smoothing and so on. It is this approach that this paper focuses on.

The idea of this kind of adaptive Bayesian learning for HMM is not a new one. By assuming that the set of vectors assigned to each prototype is modeled by a diagonal multivariate Gaussian density, of which the prototype is the mean, Ferretti and Scarci [10] used Bayesian estimation of mean vectors to build speaker-specific codebooks in an DHMM (Discrete Hidden Markov Model) framework. Originated in Brown *et al*'s work with Bayesian estimation for speaker adaptation of CDHMM (Continuous Density Hidden Markov Model) parameters in a connected digit recognizer [5], recently Lee *et al* [19] investigated various Bayesian training schemes for the speaker adaptation in isolated word recognition where the parameters of multivariate Gaussian state observation density with diagonal covariance

matrix were adapted, and the same Bayesian adaptation procedure can be easily extended to cope with the case of a multivariate Gaussian density with a full covariance matrix.[1] Later Gauvain *et al* [12] managed to extend Bayesian adaptation to handle parameters of mixture of gaussian densities with diagonal covariance matrix. They proposed to use a prior density which is the product of a Dirichlet density and gamma-normal densities. By further assuming two regularity conditions they used the EM algorithm [9] to iteratively find the mode of the posterior density. This very special assumption of regularity conditions may limit the ability of this kind of prior density to represent prior information adequately. As a matter of fact, EM algorithm needs no regularity condition.

So far, Bayesian adaptive learning in HMM training applies to only the adaptation of either the codebook in the DHMM framework or the state observation densities in CDHMM. Nothing about adaptation of the initial state distribution, the transition matrix or the state observation distribution in DHMM has been reported in the literature. However, it is recently learned from C. H. Lee that they have extended the MAP learning to all HMM parameters with general mixture Gaussian state observation densities [13]. Hence this paper will only focus on the problem of Bayesian adaptive learning for DHMM and Semi-continuous HMM (SCHMM).

The rest of the paper is organized as follows: After a brief introduction of the concept of the Bayesian point estimation in Section 2, the formulation of MAP estimates for DHMM and SCHMM are derived respectively in Section 3 and 4. In Section 5, the problem of segmental MAP estimates for HMM are discussed and a computationally efficient method of segmental quasi-Bayes estimation for SCHMM is presented. The important issue of prior density estimation is discussed in Section 6 and a simplified method of moment estimate is given. Finally the findings are summarized in Section 7.

2 Bayesian Point Estimation

In the Bayesian approach, if θ is the unknown parameter vector to be estimated from a sequence of n observations $x_1, x_2, \cdots, x_n$, it is assumed that an investigator's prior knowledge about θ can be summarized in a prior probability density function (PDF) $p(\theta)$, with $\theta \in \Omega$,

[1] In Lee *et al* [19], Eq(3.18) $\hat{\sigma}^2 = \hat{\beta}/\hat{\alpha}$ is not a MAP estimate of σ^2. It is the Bayesian point estimate of σ^2 with the quadratic loss function. The true MAP estimate of σ^2 must be $\hat{\beta}/(\hat{\alpha} - 0.5)$.

where Ω denotes an admissible region of the parameter space.[2] By the use of Bayes' theorem, this information can be combined with the sample density function $p(x_1, x_2, \cdots, x_n|\theta)$ (which is the likelihood function if viewed as a function of θ) to yield a posterior PDF $p(\theta|x_1, x_2, \cdots, x_n)$. Such a PDF can be used to make inferences about the parameter θ:

$$p(\theta|x_1, x_2, \cdots, x_n) = \frac{p(x_1, x_2, \cdots, x_n|\theta)p(\theta)}{\int_\Omega p(x_1, x_2, \cdots, x_n|\theta)p(\theta)d\theta} \tag{1}$$

Furthermore, if an investigator has a loss function which reflects the cost of an incorrect estimation, it is generally possible to obtain an estimate, say $\hat{\theta}$, which minimizes the posterior expected loss. Under a wide range of conditions, $\hat{\theta}$ will also be a function of the sample observations which minimizes the average risk. In this latter case, $\hat{\theta}$ is formally termed the Bayesian point estimator relative to the given loss function and prior PDF employed. It is well known that the mean of the posterior PDF is the Bayesian point estimator given that the loss function is quadratic while the mode of the posterior PDF is the one usually called modal or MAP (maximum *a posteriori*) estimator corresponding to the special zero-one loss function structure. Both of them are reasonable candidate of the point estimate of θ [18, 8, 28]. In particular, when the prior PDF $p(\theta)$ is constant over the parameter space Ω, the MAP estimator is the same as a classical maximum likelihood (ML) estimator.

3 MAP Estimate for Discrete HMM

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

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

$$g(\lambda) = g(\pi) \cdot g(A) \cdot g(B) \tag{2}$$

[2] In denoting the prior PDF $p(\theta)$, we do not explicitly show the parameters of the prior PDF which are assigned values by the investigator. Also note that for simplicity, in this paper we will use the convention that both the random variable and the value it may assume are denoted with the same symbol. Since it is not likely to cause confusion.

Such an independence assumption may not be unduly restrictive. If the rows of π, A and B are assumed independently distributed *a priori*, and their densities assume the form of Dirichlet distributions (sometimes called multivariate beta PDF), then $g(\lambda)$ becomes a special case of the matrix beta PDF [22]:

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

where K_c is a normalizing factor. $\{\eta_i\}, \{\eta_{ij}\}, \{\nu_{ik}\}$ are sets of positive parameters for the prior PDF of π, A, B assigned by the investigator to represent his prior knowledge of the parameters.

Assuming a prior distribution as a Dirichlet one is not without criticism [1], but it does lead to a tractable analysis and a development of subjective elicitation procedures had been reported [6, 7]. Also note that the "extended natural conjugate" prior distribution which admits non-zero correlation between the rows of A, B, and π will result in complicated formulas for the moments, etc. [22].

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

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

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

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

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

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

where the summation is taken over all possible state sequences.

Given the observation sequence $\mathbf{x}$ and the prior density $g(\lambda)$, the MAP estimate of λ can be obtained by

$$\lambda_{MAP} = arg \max_{\lambda} P(\mathbf{x}|\lambda) g(\lambda) \tag{7}$$

By viewing it as a missing data problem, as noted by Dempster et al [9], the EM (expectation-maximization) algorithm can be easily modified to produce this MAP estimate.

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

$$\log P(\mathbf{x}, \mathbf{s}|\lambda) = \log \pi_{s_1} + \sum_{t=2}^{T} \log a_{s_{t-1}s_t} + \sum_{t=1}^{T} \log b_{s_t}(x_t) \tag{8}$$

Define an auxiliary function $R(\hat{\lambda}|\lambda) = Q(\hat{\lambda}|\lambda) + \log g(\hat{\lambda})$, where $Q(\hat{\lambda}|\lambda)$ is the auxiliary function for the E-step in ML estimation.

$$\begin{aligned} Q(\hat{\lambda}|\lambda) &= E[\log P(\mathbf{x}, \mathbf{s}|\hat{\lambda})|\mathbf{x}, \lambda] && (9)\\ &= \sum_{\mathbf{s}} \{\frac{P(\mathbf{x}, \mathbf{s}|\lambda)}{P(\mathbf{x}|\lambda)} \log P(\mathbf{x}, \mathbf{s}|\hat{\lambda})\} && (10)\\ &= \sum_{i=1}^{N} e_i \log \hat{\pi}_i + \sum_{i=1}^{N}\sum_{j=1}^{N} c_{ij} \log \hat{a}_{ij} + \sum_{j=1}^{N}\sum_{k=1}^{K} d_{jk} \log \hat{b}_{jk} && (11) \end{aligned}$$

where

$$\begin{aligned} e_i &= Pr(s_1 = i|\mathbf{x}, \lambda) && (12)\\ c_{ij} &= \sum_{t=1}^{T-1} Pr(s_t = i, s_{t+1} = j|\mathbf{x}, \lambda) && (13)\\ d_{jk} &= \sum_{t:x_t \sim v_k} Pr(s_t = j, x_t \sim v_k|\mathbf{x}, \lambda) && (14) \end{aligned}$$

and these terms can be efficiently computed by using the Forward-Backward algorithm [23]. Thus,

$$R(\hat{\lambda}|\lambda) = Q(\hat{\lambda}|\lambda) + \sum_{i=1}^{N} (\eta_i - 1) \log \hat{\pi}_i + \sum_{i=1}^{N}\sum_{j=1}^{N} (\eta_{ij} - 1) \log \hat{a}_{ij} + \sum_{j=1}^{N}\sum_{k=1}^{K} (\nu_{jk} - 1) \log \hat{b}_{jk} + \log K_c \tag{15}$$

where K_c is just a function of $\{\eta_i\}$, $\{\eta_{ij}\}$, and $\{\nu_{ik}\}$, not dependent on $\hat{\lambda}$. By choosing $\hat{\lambda}$ to maximize $R(\hat{\lambda}|\lambda)$, the EM iteration for the three parameter sets π, A, B is as follows:[3]

$$\hat{\pi}_i = \frac{e_i + \eta_i - 1}{\sum_{i=1}^{N} e_i + \sum_{i=1}^{N} \eta_i - N} \qquad i = 1, 2, \cdots, N \tag{16}$$

$$\hat{a}_{ij} = \frac{c_{ij} + \eta_{ij} - 1}{\sum_{j=1}^{N} c_{ij} + \sum_{j=1}^{N} \eta_{ij} - N} \qquad i, j = 1, 2, \cdots, N \tag{17}$$

$$\hat{b}_{jk} = \frac{d_{jk} + \nu_{jk} - 1}{\sum_{k=1}^{K} d_{jk} + \sum_{k=1}^{K} \nu_{jk} - K} \qquad j = 1, 2, \cdots, N; \quad k = 1, 2, \cdots, K \tag{18}$$

[3]Strictly speaking, three conditions must be obeyed: (1) $e_i + \eta_i > 1$, (2) $c_{ij} + \eta_{ij} > 1$ and (3) $d_{jk} + \nu_{jk} > 1$. This is usually the case in practice; otherwise, these simple formulas cannot be derived.

If there are multiple independent observation sequences $\{\mathbf{x}_w\}_{w=1,\cdots,W}$, with $\mathbf{x}_w = (x_1^{(w)}, \cdots, x_{T_w}^{(w)})$, to get an MAP estimate of λ, one just maximizes $g(\lambda)\prod_{w=1}^{W} P(\mathbf{x}_w|\lambda)$, where $P(\mathbf{x}_w|\lambda)$ is as defined in equation (6). The EM auxiliary function will then become

$$R(\hat{\lambda}|\lambda) = \log g(\hat{\lambda}) + \sum_{w=1}^{W} E[\log P(\mathbf{x}_w, \mathbf{s}_w|\hat{\lambda})|\mathbf{x}_w, \lambda] \tag{19}$$

where $P(\mathbf{x}_w, \mathbf{s}_w|\hat{\lambda})$ is defined by equation (5). It is straightforward to derive the following reestimation formulas:

$$\hat{\pi}_i = \frac{\sum_{w=1}^{W} e_i^{(w)} + \eta_i - 1}{\sum_{w=1}^{W}\sum_{i=1}^{N} e_i^{(w)} + \sum_{i=1}^{N}\eta_i - N} \qquad i = 1, 2, \cdots, N \tag{20}$$

$$\hat{a}_{ij} = \frac{\sum_{w=1}^{W} c_{ij}^{(w)} + \eta_{ij} - 1}{\sum_{w=1}^{W}\sum_{j=1}^{N} c_{ij}^{(w)} + \sum_{j=1}^{N}\eta_{ij} - N} \qquad i, j = 1, 2, \cdots, N \tag{21}$$

$$\hat{b}_{jk} = \frac{\sum_{w=1}^{W} d_{jk}^{(w)} + \nu_{jk} - 1}{\sum_{w=1}^{W}\sum_{k=1}^{K} d_{jk}^{(w)} + \sum_{k=1}^{K}\nu_{jk} - K} \qquad j = 1, 2, \cdots, N; \quad k = 1, 2, \cdots, K \tag{22}$$

where $e_i^{(w)}$, $c_{ij}^{(w)}$, $d_{jk}^{(w)}$ are obtained by applying the Forward-Backward algorithm for each observation sequence $\mathbf{x}_w$.

Note that when $W \to \infty$, the MAP reestimation formulas approach the Baum-Welch ones which are used to get an approximate ML estimate. Thus an asymptotical similarity of the two estimates is demonstrated. Iterative use of these reestimation formulas will give the estimates of the HMM parameters which correspond to a local maximum of the posterior density, provided the iterative sequence is not trapped at some saddle point, in which case, a small random perturbation of λ away from the saddle point will hopefully set the EM algorithm free from the saddle point. The reader is referred to the detailed account of the convergence properties of the EM algorithm in a general setting given by Wu [29]. The choice of the initial estimates is therefore essential for finding a "good" solution and minimizing the number of EM iterations needed to attain a local maximum. One reasonable choice of the initial estimates is the mode of the prior density [4]:

$$\pi_i^{(0)} = \frac{\eta_i - 1}{\sum_{i=1}^{N}\eta_i - N} \qquad i = 1, 2, \cdots, N \tag{23}$$

$$a_{ij}^{(0)} = \frac{\eta_{ij} - 1}{\sum_{j=1}^{N}\eta_{ij} - N} \qquad i, j = 1, 2, \cdots, N \tag{24}$$

[4]Note again that the following three conditions must be obeyed: (1) $\eta_i > 1$, (2) $\eta_{ij} > 1$, and (3) $\nu_{jk} > 1$. This is usually the case in practice; otherwise, no simple formulas can be derived.

$$b_{jk}^{(0)} = \frac{\nu_{jk} - 1}{\sum_{k=1}^{K} \nu_{jk} - K} \qquad j = 1, 2, \cdots, N; \quad k = 1, 2, \cdots, K \tag{25}$$

Another choice for initial values is the mean of the prior density:

$$\pi_i^{(0)} = \frac{\eta_i}{\sum_{i=1}^{N} \eta_i} \qquad i = 1, 2, \cdots, N \tag{26}$$

$$a_{ij}^{(0)} = \frac{\eta_{ij}}{\sum_{j=1}^{N} \eta_{ij}} \qquad i, j = 1, 2, \cdots, N \tag{27}$$

$$b_{jk}^{(0)} = \frac{\nu_{jk}}{\sum_{k=1}^{K} \nu_{jk}} \qquad j = 1, 2, \cdots, N; \quad k = 1, 2, \cdots, K \tag{28}$$

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

4 MAP Estimate for Semi-continuous HMM

Semi-continuous [14] or tied mixture [4] HMM has its distinctive advantage in modeling speech for recognition. In this section, we will discuss the MAP estimate for Semi-continuous HMM (SCHMM), where the state observation densities are mixtures of Gaussian PDFs:

$$b_j(x_t) = \sum_{k=1}^{K} \omega_{jk} f(x_t | \theta_k) \tag{29}$$

$$= \sum_{k=1}^{K} \omega_{jk} N(x_t | m_k, r_k) \tag{30}$$

where $N(x|m_k, r_k)$ is the k-th normal mixand denoted by

$$N(x|m_k, r_k) \propto |r_k|^{1/2} \exp[-\frac{1}{2}(x - m_k)^t r_k (x - m_k)] \tag{31}$$

Here " $\propto$ " denotes proportionality, m_k is the D-dimensional mean vector and r_k is the $D \times D$ precision matrix (a precision matrix is defined as the inverse of the covariance matrix) [5]. These Gaussian mixture components are shared by all the states of every HMM. Each state observation density differs from another by its corresponding mixture coefficients ω_{jk}, which satisfies the constraint $\sum_{k=1}^{K} \omega_{jk} = 1$.

Thus, a SCHMM is represented by a parameter vector $\lambda = (\pi, A, \theta)$, where π is the initial state distribution, A is the transition matrix, and θ is the PDF parameter vector composed of the mixture parameters $\theta_i = \{\omega_{ik}, m_k, r_k\}_{k=1,2,\cdots,K}$ for each state i. Since the SCHMMs

[5] $|r|$ 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.

share the mixture components in state observation density, different models must be estimated simultaneously. Now consider a collection of M SCHMMs, $\Lambda = (\lambda_1, \lambda_2, \cdots, \lambda_M)$. The prior density for Λ is assumed to be:

$$g(\Lambda) = [\prod_{m=1}^{M} g(\lambda_m)] \prod_{k=1}^{K} g(m_k, r_k) \tag{32}$$

where

$$g(\lambda_m) \propto \prod_{i=1}^{N} \{[\pi_i^{(m)}]^{\eta_i^{(m)}-1} \cdot (\prod_{j=1}^{N} [a_{ij}^{(m)}]^{\eta_{ij}^{(m)}-1}) \cdot (\prod_{k=1}^{K} [\omega_{ik}^{(m)}]^{\nu_{ik}^{(m)}-1})\} \tag{33}$$

taking the form in equation (3).

If the Gaussian mixand has a full covariance matrix, then $g(m_k, r_k)$ is assumed to be a normal-Wishart density [8] of the form

$$g(m_k, r_k|\tau_k, \mu_k, \alpha_k, u_k) \propto |r_k|^{(\alpha_k - D)/2} \exp[-\frac{\tau_k}{2}(m_k - \mu_k)^t r_k (m_k - \mu_k)] \exp[-\frac{1}{2} tr(u_k r_k)] \tag{34}$$

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

On the other hand, if the Gaussian mixand has a diagonal covariance matrix, then $g(m_k, r_k)$ is assumed to be a product of normal-gamma density [8] with the form:

$$g(m_k, r_k|\tau_{kd}, \mu_{kd}, \alpha_{kd}, \beta_{kd}) \propto \prod_{d=1}^{D} r_{kd}^{(\alpha_{kd} - 1/2)} \exp[-\frac{1}{2}\tau_{kd} r_{kd} (m_{kd} - \mu_{kd})^2] \exp[-\beta_{kd} r_{kd}] \tag{35}$$

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

Let $\mathbf{x}^{(m,n)}$ denote the nth observation sequence of length $T^{(m,n)}$ associated with model m, and each model has W_m such observation sequences. Let λ_m denote the set of parameters of the m-th HMM.

Given the set of observation sequences $\{\mathbf{x}^{(m,n)}\}$ and the above prior PDF $g(\Lambda)$, the MAP estimates of Λ can be obtained by

$$\Lambda_{MAP} = arg \max_{\Lambda} \{\prod_{m=1}^{M} \prod_{n=1}^{W_m} f(\mathbf{x}^{(m,n)}|\lambda_m)\} \cdot g(\Lambda) \tag{36}$$

This can also be solved by the EM algorithm.

Define a general Q-function as

$$Q(\hat{\Lambda}|\Lambda) = \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{\mathbf{s}^{(m,n)}} \sum_{\mathbf{l}^{(m,n)}} \frac{f(\mathbf{x}^{(m,n)}, \mathbf{s}^{(m,n)}, \mathbf{l}^{(m,n)}|\lambda_m)}{f(\mathbf{x}^{(m,n)}|\lambda_m)} \log f(\mathbf{x}^{(m,n)}, \mathbf{s}^{(m,n)}, \mathbf{l}^{(m,n)}|\hat{\lambda}_m) \tag{37}$$

where $\mathbf{s}^{(m,n)}$ is the unobserved state sequence and $\mathbf{l}^{(m,n)}$ is the sequence of the unobserved mixture component labels correspond to the observation sequence $\mathbf{x}^{(m,n)}$. Furthermore,

$$f(\mathbf{x},\mathbf{s},\mathbf{l}|\lambda) \;=\; \pi_{s_1}\omega_{s_1 l_1}N(x_1|m_{l_1},r_{l_1})\prod_{t=2}^{T}\{a_{s_{t-1}s_t}\omega_{s_{t-1}s_t}N(x_t|m_{l_t},r_{l_t})\} \tag{38}$$

and

$$f(\mathbf{x}|\lambda) \;=\; \sum_{\mathbf{s}}\{\pi_{s_1}b_{s_1}(x_1)\prod_{t=2}^{T}a_{s_{t-1}s_t}b_{s_t}(x_t)\} \tag{39}$$

It is straightforward to derive that:

$$Q(\hat{\Lambda}|\Lambda) \;=\; \sum_{m=1}^{M}Q_{\hat{\pi}}^{(m)}(\hat{\pi},\lambda)+\sum_{m=1}^{M}Q_{\hat{A}}^{(m)}(\hat{A},\lambda)+\sum_{m=1}^{M}Q_{\hat{\omega}}^{(m)}(\hat{\omega},\lambda)+\sum_{k=1}^{K}Q_{\hat{\theta}_k}(\hat{\theta}_k,\Lambda) \tag{40}$$

where

$$Q_{\hat{\pi}}^{(m)}(\hat{\pi},\lambda) \;=\; \sum_{i=1}^{N}\sum_{n=1}^{W_m}\gamma_1^{(m,n)}(i)\log\hat{\pi}_i^{(m)} \tag{41}$$

$$Q_{\hat{A}}^{(m)}(\hat{A},\lambda) \;=\; \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\gamma_t^{(m,n)}(i,j)\log\hat{a}_{ij}^{(m)} \tag{42}$$

$$Q_{\hat{\omega}}^{(m)}(\hat{\omega},\lambda) \;=\; \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\zeta_t^{(m,n)}(i,k)\log\hat{\omega}_{ik}^{(m)} \tag{43}$$

$$Q_{\hat{\theta}_k}(\hat{\theta}_k,\Lambda) \;=\; \sum_{m=1}^{M}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\zeta_t^{(m,n)}(k)\log N(x_t^{(m,n)}|\hat{m}_k,\hat{r}_k) \tag{44}$$

with

$$\gamma_t^{(m,n)}(i,j) \;=\; Pr(s_t^{(m,n)}=i,s_{t+1}^{(m,n)}=j|\mathbf{x}^{(m,n)},\lambda_m)\quad 1\le t\le T^{(m,n)}-1 \tag{45}$$

$$\gamma_t^{(m,n)}(i) \;=\; Pr(s_t^{(m,n)}=i|\mathbf{x}^{(m,n)},\lambda_m)\quad 1\le t\le T^{(m,n)} \tag{46}$$

$$\zeta_t^{(m,n)}(i,k) \;=\; Pr(s_t^{(m,n)}=i,l_t^{(m,n)}=k|\mathbf{x}^{(m,n)},\lambda_m)\quad 1\le t\le T^{(m,n)} \tag{47}$$

$$\zeta_t^{(m,n)}(k) \;=\; Pr(l_t^{(m,n)}=k|\mathbf{x}^{(m,n)},\lambda_m)\quad 1\le t\le T^{(m,n)} \tag{48}$$

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

$$\zeta_t(i,k) \;=\; \gamma_t(i)\cdot\frac{\omega_{ik}N(x_t|m_k,r_k)}{\sum_{k=1}^{K}\omega_{ik}N(x_t|m_k,r_k)} \tag{49}$$

These terms can be computed efficiently by using the Forward-Backward algorithm [23].

The MAP auxiliary function is $R(\hat{\Lambda}|\Lambda) = Q(\hat{\Lambda}|\Lambda) + \log g(\hat{\Lambda})$. With the form chosen for $g(\hat{\Lambda})$ as in equation (32),

$$\begin{aligned}\log g(\hat{\Lambda}) &= \sum_{m=1}^{M}\sum_{i=1}^{N}(\eta_i^{(m)}-1)\log\hat{\pi}_i^{(m)} + \sum_{m=1}^{M}\sum_{i=1}^{N}\sum_{j=1}^{N}(\eta_{ij}^{(m)}-1)\log\hat{a}_{ij}^{(m)} \\ &+ \sum_{m=1}^{M}\sum_{i=1}^{N}\sum_{k=1}^{K}(\nu_{jk}^{(m)}-1)\log\hat{\omega}_{jk}^{(m)} + \sum_{k=1}^{K}\log g(\hat{m}_k,\hat{r}_k) + Constant \qquad (50)\end{aligned}$$

The "M-step" in EM algorithm now becomes $\max_{\hat{\Lambda}} R(\hat{\Lambda}|\Lambda)$ and the reestimation formulas for the $\{\pi_i^{(m)}\}$, $a_{ij}^{(m)}$, $\omega_{ik}^{(m)}$ can be easily derived as :

$$\hat{\pi}_i^{(m)} = \frac{\eta_i^{(m)} - 1 + \sum_{n=1}^{W_m}\gamma_1^{(m,n)}(i)}{\sum_{i=1}^{N}\eta_i^{(m)} - N + \sum_{n=1}^{W_m}\sum_{i=1}^{N}\gamma_1^{(m,n)}(i)} \qquad i = 1,2,\cdots,N \qquad (51)$$

$$\hat{a}_{ij}^{(m)} = \frac{\eta_{ij}^{(m)} - 1 + \sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\gamma_t^{(m,n)}(i,j)}{\sum_{j=1}^{N}\eta_{ij}^{(m)} - N + \sum_{n=1}^{W_m}\sum_{j=1}^{N}\sum_{t=1}^{T^{(m,n)}}\gamma_t^{(m,n)}(i,j)} \qquad i,j = 1,2,\cdots,N \qquad (52)$$

$$\hat{\omega}_{ik}^{(m)} = \frac{\nu_{ik}^{(m)} - 1 + \sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\zeta_t^{(m,n)}(i,k)}{\sum_{k=1}^{K}\nu_{ik}^{(m)} - K + \sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\sum_{k=1}^{K}\zeta_t^{(m,n)}(i,k)} \qquad i = 1,2,\cdots,N;\ \ k = 1,2,\cdots,K \qquad (53)$$

The reestimation formulas for $\{\hat{m}_k\}$, $\{\hat{r}_k\}$ can be derived by maximizing

$$Q_{\hat{\theta}_k}(\hat{\theta}_k,\Lambda) + \log g(\hat{m}_k,\hat{r}_k) \qquad (54)$$

which leads to solving the following equations:

$$\sum_{m=1}^{M}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\zeta_t^{(m,n)}(k)\frac{\partial}{\partial\hat{m}_k}\log N(x_t^{(m,n)}|\hat{m}_k,\hat{r}_k) + \frac{\partial}{\partial\hat{m}_k}\log g(\hat{m}_k,\hat{r}_k) = 0 \qquad (55)$$

and

$$\sum_{m=1}^{M}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}}\zeta_t^{(m,n)}(k)\frac{\partial}{\partial\hat{r}_k}\log N(x_t^{(m,n)}|\hat{m}_k,\hat{r}_k) + \frac{\partial}{\partial\hat{r}_k}\log g(\hat{m}_k,\hat{r}_k) = 0 \qquad (56)$$

4.1 Full Covariance Matrix Case

Notice that when r_k is a full covariance matrix,

$$\frac{\partial}{\partial\hat{m}_k}\log N(x_t^{(m,n)}|\hat{m}_k,\hat{r}_k) = \hat{r}_k(x_t^{(m,n)} - \hat{m}_k) \qquad (57)$$

$$\frac{\partial}{\partial\hat{r}_k}\log N(x_t^{(m,n)}|\hat{m}_k,\hat{r}_k) = \frac{1}{2}[\hat{r}_k^{-1} - (x_t^{(m,n)} - \hat{m}_k)(x_t^{(m,n)} - \hat{m}_k)^t] \qquad (58)$$

and

$$\frac{\partial}{\partial \hat{m}_k} \log g(\hat{m}_k, \hat{r}_k) \;=\; -\tau_k \hat{r}_k (\hat{m}_k - \mu_k) \tag{59}$$

$$\frac{\partial}{\partial \hat{r}_k} \log g(\hat{m}_k, \hat{r}_k) \;=\; \frac{\alpha_k - D}{2} \hat{r}_k^{-1} - \frac{\tau_k}{2} (\hat{m}_k - \mu_k)(\hat{m}_k - \mu_k)^t - \frac{1}{2} u_k \tag{60}$$

Substitute these terms into equation (55) and (56), the reestimation formulas for $\hat{m}_k$, $\hat{r}_k$ can be easily obtained as:

$$\hat{m}_k \;=\; \frac{\tau_k \mu_k + \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \zeta_t^{(m,n)}(k) x_t^{(m,n)}}{\tau_k + \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \zeta_t^{(m,n)}(k)} \tag{61}$$

$$\hat{r}_k^{-1} = \frac{u_k + \tau_k (\hat{m}_k - \mu_k)(\hat{m}_k - \mu_k)^t + \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \zeta_t^{(m,n)}(k) (x_t^{(m,n)} - \hat{m}_k)(x_t^{(m,n)} - \hat{m}_k)^t}{\alpha_k - D + \sum_{m=1}^{M} \sum_{n=1}^{W_m} \sum_{t=1}^{T^{(m,n)}} \zeta_t^{(m,n)}(k)} \tag{62}$$

These two equations together with equations (51) to (53) constitute the MAP reestimation formulas for Λ. The initial estimate can be chosen as the mode of the prior PDF $g(\Lambda)$: $\{\pi_i^{(m)}\}$, $\{a_{ij}^{(m)}\}$, $\{\omega_{ik}^{(m)}\}$ have the same form as equation (23) $\sim$ (25) in the case of DHMM, and

$$m_k \;=\; \mu_k \tag{63}$$

$$r_k \;=\; (\alpha_k - D) u_k^{-1} \tag{64}$$

Another choice is the mean of the prior PDF $g(\Lambda)$: $\{\pi_i^{(m)}\}$, $\{a_{ij}^{(m)}\}$, $\{\omega_{ik}^{(m)}\}$ also have the same form as equation (26) $\sim$ (28) and

$$m_k \;=\; \mu_k \tag{65}$$

$$r_k \;=\; \alpha_k u_k^{-1} \tag{66}$$

4.2 Diagonal Covariance Matrix Case

When r_k is a diagonal covariance matrix,

$$\frac{\partial}{\partial \hat{m}_{kd}} \log N(x_t^{(m,n)} | \hat{m}_k, \hat{r}_k) \;=\; \hat{r}_{kd} (x_{td}^{(m,n)} - \hat{m}_{kd}) \tag{67}$$

$$\frac{\partial}{\partial \hat{r}_{kd}} \log N(x_t^{(m,n)} | \hat{m}_k, \hat{r}_k) \;=\; \frac{1}{2} [\hat{r}_{kd}^{-1} - (x_{td}^{(m,n)} - \hat{m}_{kd})^2] \tag{68}$$

and

$$\frac{\partial}{\partial \hat{m}_{kd}} \log g(\hat{m}_k, \hat{r}_k) \;=\; -\tau_{kd} \hat{r}_{kd} (\hat{m}_{kd} - \mu_{kd}) \tag{69}$$

$$\frac{\partial}{\partial \hat{r}_{kd}} \log g(\hat{m}_k, \hat{r}_k) \;=\; (\alpha_{kd} - \frac{1}{2})\hat{r}_{kd}^{-1} - \frac{\tau_{kd}}{2}(\hat{m}_{kd} - \mu_{kd})^2 - \beta_{kd} \tag{70}$$

Substitute these terms into equation (55) and (56), the reestimation formulas for $\hat{m}_{kd}$, $\hat{r}_{kd}$ can be easily derived as:

$$\hat{m}_{kd} \;=\; \frac{\tau_{kd}\mu_{kd} + \sum_{m=1}^{M}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}} \zeta_t^{(m,n)}(k) x_{td}^{(m,n)}}{\tau_{kd} + \sum_{m=1}^{M}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}} \zeta_t^{(m,n)}(k)} \tag{71}$$

$$\hat{r}_{kd}^{-1} \;=\; \frac{2\beta_{kd} + \tau_{kd}(\hat{m}_{kd} - \mu_{kd})^2 + \sum_{m=1}^{M}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}} \zeta_t^{(m,n)}(k)(x_{td}^{(m,n)} - \hat{m}_{kd})^2}{2\alpha_{kd} - 1 + \sum_{m=1}^{M}\sum_{n=1}^{W_m}\sum_{t=1}^{T^{(m,n)}} \zeta_t^{(m,n)}(k)} \tag{72}$$

The initial estimates of m_{kd} and r_{kd} can be chosen as the mode of the prior PDF $g(\Lambda)$:

$$m_{kd} \;=\; \mu_{kd} \tag{73}$$

$$r_{kd} \;=\; (\alpha_{kd} - \frac{1}{2})/\beta_{kd} \tag{74}$$

or the mean of the prior PDF $g(\Lambda)$:

$$m_{kd} \;=\; \mu_{kd} \tag{75}$$

$$r_{kd} \;=\; \alpha_{kd}/\beta_{kd} \tag{76}$$

5 Segmental MAP Estimates for HMM

Analogous to the segmental k-means algorithm [24, 17], a similar optimization criterion can be considered for the MAP estimate of HMM. Instead of maximizing $g(\lambda|\mathbf{x})$, $g(\lambda, \mathbf{s}|\mathbf{x})$, the joint posterior density of parameters λ and state sequence $\mathbf{s}$ is maximized. The estimation procedure becomes:

$$\tilde{\lambda} \;=\; arg\max_{\lambda}\max_{\mathbf{s}} g(\lambda, \mathbf{s}|\mathbf{x}) \;=\; arg\max_{\lambda}\max_{\mathbf{s}} f(\mathbf{x}, \mathbf{s}|\lambda)g(\lambda) \tag{77}$$

Here $\tilde{\lambda}$ is called the segmental MAP estimate of λ [19]. Just as it is the case with the segmental k-means algorithm, it is straightforward to prove that starting with any estimate $\lambda^{(p)}$, 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. $g(\lambda^{(p+1)}, \mathbf{s}^{(p+1)}|\mathbf{x}) \geq g(\lambda^{(p)}, \mathbf{s}^{(p)}|\mathbf{x})$ with

$$\mathbf{s}^{(p)} \;=\; arg\max_{\mathbf{s}} f(\mathbf{x}, \mathbf{s}|\lambda^{(p)}) \tag{78}$$

$$\lambda^{(p+1)} \;=\; arg\max_{\lambda} f(\mathbf{x}, \mathbf{s}^{(p)}|\lambda)g(\lambda) \tag{79}$$

The most likely state sequence $\mathbf{s}^{(p)}$ is decoded by the Viterbi algorithm [11]. If the maximization over λ in (79) has no closed form solution, it can be replaced by any hill climbing procedure which replaces $\lambda^{(p)}$ by $\lambda^{(p+1)}$ subject to the constraint that $f(\mathbf{x}, \mathbf{s}^{(p)}|\lambda^{(p+1)})g(\lambda^{(p+1)}) \geq f(\mathbf{x}, \mathbf{s}^{(p)}|\lambda^{(p)})g(\lambda^{(p)})$.

5.1 Segmental MAP Estimate for DHMM

By applying the Viterbi algorithm to the training data, apart from the most likely state sequences, the sets of observations associated with each HMM state are also available. Let $n_i^{(1)}$ denote the numbers of observations in state i at time $t = 1$, and n_{ij} be the transition count from state i to state j in the most likely state sequences. Furthermore, let f_{jk} denote the count of observing symbol v_k in state j. It is straight forward to show that the reestimation formulas in equation (16) to (18) are the closed form solution of (79) by replacing the e_i by $n_i^{(1)}$, c_{ij} by n_{ij} and d_{jk} by f_{jk}.

5.2 Segmental MAP Estimate for SCHMM

The reestimation formulas for $\{\pi_i\}$ and $\{a_{ij}\}$ are the same as that in DHMM. By replacing $\zeta_t^{(m,n)}(i,k)$ in equation (49) by

$$\zeta_t^{(m,n)}(i,k) \;=\; \delta(s_t^{(m,n)} - i) \cdot \frac{\omega_{ik}^{(m)} N(x_t^{(m,n)}|m_k, r_k)}{\sum_{k=1}^{K} \omega_{ik}^{(m)} N(x_t^{(m,n)}|m_k, r_k)} \tag{80}$$

where $\mathbf{s}^{(m,n)}$ is the most likely state sequence corresponding to observation sequence $\mathbf{x}^{(m,n)}$, and $\delta(\cdot)$ denotes the Kronecker delta function. The reestimation formulas in equation (53), (61), (62) and (71), (72) still hold.

Note that within an outer loop of iteration to update the HMM parameters, by making a single adjustment, $\{\omega_{ik}^{(m)}\}$, $\{m_k\}$, $\{r_k\}$ can be updated synchronously with the update of $\{\pi_i^{(m,n)}\}$, $\{a_{ij}^{(m,n)}\}$. Another extreme alternative which may need less global (outer) iterations is that $\{\omega_{ik}^{(m)}\}$, and/or $\{m_k\}$, $\{r_k\}$ are first updated by an inner loop of iterative adjustments to their "optimal" values (which is usually very time consuming) based on the current labeling of the training data before $\{\pi_i^{(m,n)}\}$, $\{a_{ij}^{(m,n)}\}$ are updated to get the new labeling of the training data. A compromise can be updating $\{\omega_{ik}^{(m)}\}$ (or simultaneously

$\{m_k\}$, $\{r_k\}$) a predetermined number of times before updating the remaining parameters. The optimal scheme that allows the problem to be solved in the shortest time possible is data dependent. It is also possible to use the approximate solution for these parameters as discussed in the next subsection.

5.3 Segmental Quasi-Bayes Estimate for SCHMM

In SCHMMs, all states of all HMMs share the same mixture components, so it is reasonable to assume that these mixture components are fixed and need not be adapted in the adaptive process. By applying the Viterbi algorithm to the training data, the sets of observations associated with each HMM state are available. So the updating formula for $\{\omega_{ik}^{(m)}\}$ correspond to the maximization in equation (79) can be derived by solving the following general Bayesian estimation problem for finite mixture distribution.

Given a sequence of observations $x_1, x_2, \cdots, x_n$, conditional on $\omega = (\omega_1, \omega_2, \cdots, \omega_K)$ and density functions $f_1, f_2, \cdots, f_K$, each x_n is assumed independent with probability density:

$$p(x_n|\omega) \;=\; \sum_{i=1}^{K} \omega_i f_i(x_n) \tag{81}$$

where the ω_i's are unknown, non-negative and summed to unity while the f_i are known. Assuming that the prior density for ω has the form of a Dirichlet density

$$p(\omega) \;=\; D(\omega|\nu_1^{(0)}, \nu_2^{(0)}, \cdots, \nu_K^{(0)}) \;\propto\; \prod_{i=1}^{K} \omega_i^{\nu_i^{(0)}-1} \tag{82}$$

where $\nu_i^{(0)} \;\geq\; 0, i = 1, 2, \cdots, K$.

After observing x_1, we obtain

$$p(\omega|x_1) \;=\; \sum_{i=1}^{K} p_i(x_1) D(\omega|\nu_1^{(0)} + \delta_{i1}, \cdots, \nu_K^{(0)} + \delta_{iK}) \tag{83}$$

where

$$p_i(x_1) \;=\; \frac{f_i(x_1)\nu_i^{(0)}}{\sum_{i=1}^{K} f_i(x_1)\nu_i^{(0)}} \tag{84}$$

and

$$\delta_{ij} \;=\; \begin{cases} 1 & \text{if i=j} \\ 0 & \text{otherwise} \end{cases}$$

Many well-known approximate Bayesian learning procedures to solve this problem arise from approximating the RHS of (83) by

$$p(\omega|x_1) \approx D(\omega|\nu_1^{(0)} + \hat{\delta}_{11}, \cdots, \nu_K^{(0)} + \hat{\delta}_{1K}) \tag{85}$$

The $\hat{\delta}_{ij}$'s take values according to some specified method. Proceeding in this way, the necessary computation could be kept within reasonable bounds.

In the quasi-Bayes procedure proposed by Smith and Makov [27], it is suggested that $\hat{\delta}_{1i}$ be replaced by $p_i(x_1)$, and so

$$p(\omega|x_1) \approx D(\omega|\nu_1^{(1)}, \cdots, \nu_K^{(1)}) \tag{86}$$

where $\nu_i^{(1)} = \nu_i^{(0)} + p_i(x_1)$.

Then, subsequent updating takes place entirely within the Dirichlet family of distributions: $p(\omega|x_1, x_2, \cdots, x_n)$ is Dirichlet with parameters $\nu_i^{(n)} = \nu_i^{(n-1)} + p_i(x_n)$, where $\nu_i^{(n-1)}$ are parameters of $p(\omega|x_1, x_2, \cdots, x_{n-1})$, and

$$p_i(x_n) = \frac{f_i(x_n)\nu_i^{(n-1)}}{\sum_{i=1}^{K} f_i(x_n)\nu_i^{(n-1)}} \tag{87}$$

In the sense of the approximate posterior distribution with mean identical to that of the true distribution, the convergence properties were established in [27].

It is easily verified from the well-known properties of the Dirichlet distribution that the (quasi-) posterior mean for ω_i, after observing $x_1, x_2, \cdots, x_n$ is given by

$$\hat{\omega}_i^{(n)} = \frac{\nu_i^{(n)}}{\nu_0 + n} \tag{88}$$

and the mode of the approximate posterior density is

$$\hat{\omega}_i^{(n)} = \frac{\nu_i^{(n)} - 1}{\nu_0 + n - K} \tag{89}$$

where $\nu_0 = \nu_1^{(0)} + \nu_2^{(0)} + \cdots + \nu_K^{(0)}$. Both (88) and (89) can serve as the updating formula for mixture coefficients in the segmental quasi-Bayes learning for SCHMMs. Note that because such update is an approximation, the monotonic increasing property of the objective function will not be guaranteed, but it is believed that this scheme will lead to a reasonable estimate of the parameters for SCHMMs. Also note that the results of above quasi-Bayes method depend on the order of presentation of the x_i's. A natural choice is to present the x_i's in the order of their appearance in the training speech data.

6 Estimation of the Parameters for Prior Distribution

In the previous Sections it was assumed that the prior density $g(\lambda)$ is a member of a preassigned family of prior distributions. In pure Bayesian approach, the parameter vector φ of this family of PDFs $\{g(\cdot|\varphi)\}$ is also assumed known based on a subjective knowledge about λ. In reality, it is difficult to possess complete knowledge of the prior distribution. An attractive compromise between the classical non-Bayesian approach which uses no prior information and the full Bayesian one is to adopt the Empirical Bayes (EB) approach [25, 26, 21]. Here we use a somewhat broader interpretation of the term "empirical Bayes" than what was implied by Robbins's original definition. φ is replaced by any estimate derived from the previous observed data. Then the previous data and current data are linked in the form of a two-stage sampling scheme by a common prior PDF $g(\lambda)$ of the unknown parameters λ.

Let $\mathbf{x}$ denote the current observation set to be used to adaptively estimate λ. At the time of making the current observation there are available past observation sets $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ obtained with independent past realizations $\lambda_1, \lambda_2, \cdots, \lambda_n$. The words "current" and "past" are not necessarily taken in a strictly temporal sense. Usually $\lambda_1, \lambda_2, \cdots, \lambda_n$ are not directly observed, but they have a common prior PDF $\{g(\cdot|\varphi)\}$. The hyperparameter φ can be obtained by

$$\max_{\varphi} f(\mathbf{X}|\varphi) \;=\; \int_{\Lambda} f(\mathbf{X}|\Lambda) g(\Lambda|\varphi) d\Lambda \tag{90}$$

where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$, $\Lambda = (\lambda_1, \lambda_2, \cdots, \lambda_n)$, $f(\mathbf{X}|\Lambda) = \prod_{i=1}^{n} f(\mathbf{x}_i|\lambda_i)$ and $g(\Lambda|\varphi) = \prod_{i=1}^{n} g(\lambda_i|\varphi)$

However, the maximum likelihood estimation above bases on the marginal density $f(\mathbf{X}|\varphi)$ and is difficult to compute. To simplify the problem, we can use a modified likelihood approach [21], where the likelihood function of the unknown Λ are defined as the joint probability of $(\mathbf{X}, \Lambda)$ in the EB scheme, i.e. the likelihood function is

$$L(\mathbf{X}, \Lambda) \;=\; \prod_{i=1}^{n} f(\mathbf{x}_i|\lambda_i) g(\lambda_i|\varphi) \tag{91}$$

This likelihood function is then maximized with respect to Λ and φ.

Note that $L(\mathbf{X}, \Lambda)$ is not a likelihood function in the usual sense of the word since Λ is unobservable random variables. More research seems to be necessary to justify the use of this approach. Apart from its justification, under the current assumptions of the form of

prior PDF $g(\cdot|\varphi)$, getting the maximum (modified) likelihood estimates of Λ and φ is not trivial. To further simplify the problem, the method of moment can be used to estimate φ.

One may use the observation sets $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ to estimate the corresponding HMMs $\hat{\lambda}_1, \hat{\lambda}_2, \cdots, \hat{\lambda}_n$ with the classical Baum-Welch or segmental k-means algorithm, and then pretend to view $\hat{\lambda}_i$ as the observations with density $g(\lambda)$. In the case of DHMM where $g(\lambda)$ is assumed to have the form of equation (3), i.e. a matrix beta PDF, with the properties of the moments for matrix beta PDF, we have

$$E(\pi_i) = \frac{\eta_i}{\sum_{i=1}^{N}\eta_i} \tag{92}$$

and

$$Var(\pi_i) = \frac{\eta_i(\sum_{i=1}^{N}\eta_i - \eta_i)}{(\sum_{i=1}^{N}\eta_i)^2(\sum_{i=1}^{N}\eta_i + 1)} \tag{93}$$

$$= \frac{E(\pi_i)[1 - E(\pi_i)]}{\sum_{i=1}^{N}\eta_i + 1} \tag{94}$$

Then we have

$$\eta_i = E(\pi_i)\{\frac{E(\pi_i)[1 - E(\pi_i)]}{Var(\pi_i)} - 1\} \tag{95}$$

Similarly for η_{ij} and ν_{ik} we have

$$\eta_{ij} = E(a_{ij})\{\frac{E(a_{ij})[1 - E(a_{ij})]}{Var(a_{ij})} - 1\} \tag{96}$$

$$\nu_{ik} = E(b_{ik})\{\frac{E(b_{ik})[1 - E(b_{ik})]}{Var(b_{ik})} - 1\} \tag{97}$$

Replacing $E(\pi_i)$, $Var(\pi_i)$, $E(a_{ij})$, $Var(a_{ij})$, $E(b_{ik})$, $Var(b_{ik})$ by their corresponding sample moments with $\hat{\lambda}_1, \hat{\lambda}_2, \cdots, \hat{\lambda}_n$, the moment estimates of η_i, η_{ij}, ν_{ik} can be obtained.

In the case of SCHMM, the moment estimates of η_i, η_{ij}, ν_{ik} have the same forms as their counterparts in DHMM.

When the Gaussian mixand has diagonal covariance matrix, the prior density $g(m_k, r_k)$ has the form of equation (35). Note that

$$E(r_{kd}) = \frac{\alpha_{kd}}{\beta_{kd}} \tag{98}$$

$$Var(r_{kd}) = \frac{\alpha_{kd}}{\beta_{kd}^2} \tag{99}$$

So

$$\alpha_{kd} = \frac{[E(r_{kd})]^2}{Var(r_{kd})} \tag{100}$$

$$\beta_{kd} = \frac{E(r_{kd})}{Var(r_{kd})} \tag{101}$$

Furthermore, note that:

$$E(m_{kd}) = \mu_{kd} \tag{102}$$

$$Var(m_{kd}) = \frac{\beta_{kd}}{\tau_{kd}(\alpha_{kd}-1)} \tag{103}$$

Then

$$\mu_{kd} = E(m_{kd}) \tag{104}$$

$$\tau_{kd} = \frac{\beta_{kd}}{Var(m_{kd})(\alpha_{kd}-1)} \tag{105}$$

Also substituting the sample moments of r_{kd} and m_{kd} into the above equations, the moment estimates of α_{kd}, β_{kd}, μ_{kd}, τ_{kd} are obtained.

For the full covariance matrix case, the prior density $g(m_k, r_k)$ has the form of equation (34). It is more difficult to write down a suitable number of estimating equations for the moment estimates of τ_k, α_k, μ_k, and u_k. If one considers a more restrictive prior density family by further assuming

$$\tau_k = \alpha_k = \sum_{m=1}^{M}\sum_{i=1}^{N}\hat{\nu}_{ik}^{(m)} \tag{106}$$

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

$$\mu_k = E(m_k) \tag{107}$$

$$u_k^{-1} = \alpha_k^{-1}E(r_k) \tag{108}$$

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

When enough training data are available, the above method of moment will lead to a reasonable estimate of hyperparameters φ. This estimate may be improved by the following iterative scheme: starting with an initial estimate $\varphi^{(m)}$, get the MAP estimates $\hat{\lambda}_1, \hat{\lambda}_2, \cdots, \hat{\lambda}_n$ from $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ with any methods presented in the previous sections; and then an improved $\varphi^{(m+1)}$ can be obtained by using the above method of moment.

The physical meaning of the prior density $g(\lambda|\varphi)$ is application dependent. For example, in a speaker adaptation problem, $g(\lambda|\varphi)$ may be used to represent the information of the variability of a certain model among the different speakers. So the training data for estimating φ can be divided into different sets correspond to different speakers or speaker

groups. In another kind of application, for example, to build the context-dependent models from context-independent model, the prior density $g(\lambda|\varphi)$ will represent the variability of λ caused by the different context. So the training data will be divided into sets according to the context information. Further applications of this kind of Bayesian learning method to speech recognition can be found in [12]. Note that the prior knowledge represented by $g(\lambda|\varphi)$ does not include those deterministic ones. For example, in left-right HMMs, some parameters are known and fixed, and $g(\lambda|\varphi)$ will not include them.

The estimation of hyperparameters φ is still an open problem. Further research is needed. This is a radical problem in order to make this kind of Bayesian learning method really applicable to adaptive training of HMMs.

7 Conclusion

In this paper a theoretical framework for Bayesian adaptive learning of discrete HMM and semi-continuous one with Gaussian mixture state observation densities is presented. Corresponding to the well-known Baum-Welch and segmental k-means algorithms for training HMM, formulations of MAP and segmental MAP estimation of HMM parameters are developed. Furthermore, a computationally efficient method of the segmental quasi-Bayes estimation for semi-continuous HMM is also presented. The important issue of prior density estimation is discussed and a simplified method of moment estimate has been given. The method proposed in this paper will be applicable to some problems in HMM training for speech recognition such as sequential or batch training, model adaptation, and parameter smoothing, etc.

References

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

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

[3] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occuring in the statistical analysis of probabilistic function of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[4] J. R. Bellegarda and D. Nahamoo. Tied mixture continuous parameter modeling for speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(12):2033–2045, 1990.

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

[6] K. M. Chaloner and G. T. Duncan. Assessment of a beta prior distribution: Pm elicitation. *The Statistician, Journal of the Institute of Statisticians*, 32:174–180, 1983.

[7] K. M. Chaloner and G. T. Duncan. Some properties of the dirichlet-multinomial distribution and its use in prior elicitation. *Communications in Statistics: Theory and Methods*, 16:511–523, 1987.

[8] Morris H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill Book Company, 1970.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Royal Statistical Society, Journal, Series B*, 39(1):1–38, 1977.

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

[11] G. D. Forney. The Viterbi algorithm. *Proceedings of the IEEE*, 77(2):257–286, March 1989.

[12] Jean-Luc Gauvain and C. H. Lee. Bayesian learning for hidden Markov model with Gaussian mixture state observation densities. *Speech Communication*, 11:205–213, 1992.

[13] Jean-Luc Gauvain and C. H. Lee. MAP estimation of continuous density HMM: Theory and applications. In *Proc. DARPA Speech and Natural Language Workshop*, pages 185–190, Feb. 1992.

[14] X. D. Huang and M. A. Jack. Semicontinuous hidden Markov models for speech signals. *Computer Speech and Language*, 3:239–251, 1989.

[15] B. H. Juang. Maximum-likelihood estimation of mixture multivariate stochastic observations of Markov chains. *AT&T Technical Journal*, 64(6):1235–1249, 1985.

[16] B. H. Juang, Stephen E. Levinson, and M. M. Sondhi. Maximum likelihood estimation for multivariate mixture observations of Markov chains. *IEEE Transactions on Information Theory*, IT-32(2):307–309, March 1986.

[17] B. H. Juang and L. R. Rabiner. The segmental k-means algorithm for estimating parameters of hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(9):1639–1641, September 1990.

[18] Irving H. LaValle. *An Introduction to Probability, Decision, and Inference.* Holt, Rinehart and Winston, Inc., 1970.

[19] 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 Transactions on Signal Processing*, 39(4):806–814, April 1991.

[20] L. R. Liporace. Maximum likelihood estimation for multivariate observations of Markov sources. *IEEE Transactions on Information Theory*, IT-28:729–734, 1982.

[21] J. S. Maritz and T. Lwin. *Empirical Bayes Methods (Second Edition).* Chapman and Hall, 1989.

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

[23] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 61:268–278, 1973.

[24] L. R. Rabiner, J. G. Wilpon, and B. H. Juang. A segmental k-means training procedure for connected word recognition. *AT&T Technical Journal*, 65(3):21–31, 1986.

[25] H. Robbins. An empirical Bayes approach to statistics. In *Proc. Third Berkeley Symposium on Math. Statist. and Prob.*, pages I–157–164, 1955.

[26] H. Robbins. The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.*, 35:1–20, 1964.

[27] A. F. M. Smith and U. E. Makov. A quasi-Bayes sequential procedure for mixtures. *Royal Statistical Society, Journal, Series B*, 40(1):106–112, 1978.

[28] Robert L. Winkler. *Introduction to Bayesian Inference and Decision.* Holt, Rinehart and Winston, Inc., 1972.

[29] C. F. Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.