ELSEVIER

Speech Communication 28 (1999) 313–326

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Improving Viterbi Bayesian predictive classification via sequential bayesian learning in robust speech recognition [1]

Hui Jiang [a,b,2], Keikichi Hirose [a,*], Qiang Huo [c,3]

[a] *Department of Information and Communication Engineering, The University of Tokyo, Hongo 7-3-1, Bunkyu-ku, Tokyo 113-8656, Japan*

[b] *Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada*

[c] *Department of Computer Science and Information Systems, The University of Hong Kong, Pokfulam Road, Hong Kong*

Received 3 September 1998; received in revised form 20 January 1999; accepted 2 April 1999

Abstract

In this paper, we extend our proposed Viterbi Bayesian predictive classification (VBPC) algorithm to a new class of prior probability density function (pdf), namely a family of natural conjugate prior pdf's of the *complete-data* density in continuous density hidden Markov model (CDHMM) and their mixtures. In this way, we can on-line adapt the prior pdf via a sequential Bayesian learning algorithm when some new data are available, so that the performance of VBPC can be continuously improved. Moreover, we also study a sequential Bayesian learning strategy for CDHMM based on a finite mixture approximation of its prior/posterior density which attempts to derive a more accurate prior pdf to describe the unknown mismatches. The experimental results on a speaker-independent recognition task of isolated Japanese digits confirm the viability and the usefulness of the proposed method.

Keywords: Bayesian predictive classification (BPC); Viterbi BPC (VBPC); Sequential Bayesian learning; Robust speech recognition; Natural conjugate prior

Notations

W	a speech unit (a word)
$\boldsymbol{X}$	acoustic observation
$\boldsymbol{X}^{(n)}$	n independent samples of acoustic observations
Λ	CDHMM parameter
φ	hyperparameter of CDHMM
$\boldsymbol{s}$	state sequence in CDHMM
$\boldsymbol{l}$	mixture component label sequence in CDHMM
ι	a path in CDHMM

* Corresponding author. Tel.: +81-3-5841-6667; fax: +81-3-5841-6648; e-mail: hirose@gavo.t.u-tokyo.ac.jp

[1] This paper is based on a communication presented at ICASSP'98, and has been recommended by the Editorial Board of Speech Communication.

[2] E-mail: hjiang@crg3.uwaterloo.ca

[3] E-mail: qhuo@csis.hku.hk

0167-6393/99/$ – see front matter
PII: S0167-6393(99)00018-7

Υ	path space in CDHMM
$f(\cdot\|\cdot)$	likelihood function of CDHMM
$p(\cdot\|\cdot)$	probability density function (pdf)
$\tilde{p}(\cdot\|\cdot)$	predictive pdf
$\arg\max^{(M)}$	the operation to choose the M largest items
$\Xi^{(M)}$	the set consisting of the M largest terms

1. Introduction

Recently, the topic of robust automatic speech recognition (ASR) has been attracting increasingly more research efforts in speech community (e.g., see recent reviews in (Furui, 1997; Lee, 1998)). From the modeling point of view, an ASR procedure is described as robust if it is not very sensitive to the departure from the assumptions on which it depends, such as modeling inaccuracy, mismatch between training and testing conditions, etc. From the application point of view, robust speech recognition refers to the problem to design an automatic speech recognizer which works well for different tasks and speakers under unexpected and/or adverse conditions. Especially, how to maintain the recognizer's performance under various mismatches between training and testing conditions has recently become one of the hottest topics in robust speech recognition.

In the past few years, we have been investigating a so-called *Bayesian predictive classification* (BPC) approach to deal with various types of unknown mismatches under a general theoretical framework for CDHMM (Gaussian mixture continuous density hidden Markov model) based robust speech recognition (Huo and Lee, 1997c; Huo et al., 1997; Jiang et al., 1997, 1999). In this paper, we extend our previously proposed Viterbi BPC (VBPC) algorithm (Jiang et al., 1997, 1999) to a new class of prior probability density function (pdf), namely a family of natural conjugate prior pdf's of the complete-data density of CDHMM and their mixtures. As shown in Fig. 1, equipped with the capability of sequential Bayesian learning, we can on-line adapt the prior pdf's when some new adaptation/test data become available, so that the performance of VBPC can be improved continuously. Moreover, we also study a new sequential Bayesian learning strategy for CDHMM based on a finite mixture approximation of its prior/posterior density, by which we attempt to derive a more accurate prior pdf to describe the unknown mismatches under the VBPC framework. The proposed methods have been examined in a speaker-independent recognition task of isolated Japanese digits to deal with two types of mismatch between training and testing conditions: (i) additive white Gaussian noise caused mismatch, (ii) cross-gender mismatch. The experimental results confirm the viability and the usefulness of the algorithms.

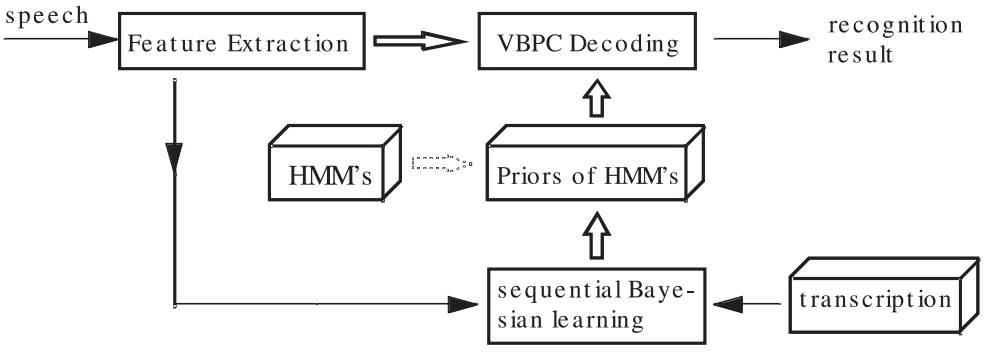

Fig. 1. A block diagram of VBPC decoder equipped with on-line Bayesian learning (supervised mode).

The remainder of the paper is organized as follows. In Section 2, after a brief introduction of the VBPC decision rule, we extend the VBPC formulation to the natural conjugate prior pdf's of the complete-data density of CDHMM. In Section 3, we introduce the basic principle of sequential Bayesian learning for CDHMM and show how to combine VBPC with a simple Bayesian learning strategy, namely segmental Bayesian learning. In Sections 4 and 5, we propose a novel Bayesian learning method for CDHMM using a finite mixture approximation of the CDHMM's true prior/posterior pdf, and develop an N-Best based implementation strategy to practically perform the above-mentioned Bayesian learning. In Section 6, we report experimental results along with some discussions. Finally, we summarize our findings in Section 7.

2. Viterbi Bayesian predictive classification

Given a speech unit (refer to as *word* heretofore) W and the associated acoustic observation $\boldsymbol{X}$, we model each word W with a CDHMM. Assuming that the model is accurate enough for $\boldsymbol{X}$ and no mismatch exists between training and testing conditions, the prior pdf of model parameter Λ can be viewed as a delta function centered at the true model parameter. An optimal speech recognizer can be achieved by optimal MAP (maximum a posteriori) decision rule. However, in case the mismatch exists between training and testing conditions, some extra uncertainty will be involved into the decision procedure due to unknown mismatch. The prior pdf of model parameter Λ is inflated to an unknown function $p(\Lambda|\varphi, W)$ with hyperparameter φ. Obviously, $p(\Lambda|\varphi, W)$ represents the prior knowledge about the involved mismatch and the interaction between the mismatch and model parameter Λ related to each word W. Under these assumptions, an optimal decision rule for robust speech recognition which achieves an expected minimum word recognition error rate in the mismatched situation is based on the following BPC decoding (Huo and Lee, 1997c; Huo et al., 1997):

$$\hat{W} = \arg\max_{W} \tilde{p}(W|\boldsymbol{X}) = \arg\max_{W} \tilde{p}(\boldsymbol{X}, W) = \arg\max_{W} \tilde{p}(\boldsymbol{X}|\varphi, W) \cdot p(W), \tag{1}$$

with

$$\tilde{p}(\boldsymbol{X}|\varphi, W) = \int f(\boldsymbol{X}|\Lambda, W) \cdot p(\Lambda|\varphi, W)\,\mathrm{d}\Lambda = \sum_{\boldsymbol{s},\boldsymbol{l}} \int f(\boldsymbol{X}, \boldsymbol{s}, \boldsymbol{l}|\Lambda, W) \cdot p(\Lambda|\varphi, W)\,\mathrm{d}\Lambda, \tag{2}$$

where $\tilde{p}(\boldsymbol{X}|\varphi, W)$ is called the predictive pdf of the observation $\boldsymbol{X}$ given the word W, $f(\boldsymbol{X}|\Lambda, W)$ is likelihood function of CDHMM, and $\boldsymbol{s}$ and $\boldsymbol{l}$ denote the unobserved state sequence and the associated sequence of the unobserved mixture component labels, respectively.

However, due to the nature of the *missing data* problem in HMM formulation (see related discussions in (Huo and Lee, 1997a,c)), it is not easy to compute the true *predictive density* $\tilde{p}(\boldsymbol{X}|\varphi, W)$ in the CDHMM case. One feasible way is to compute the predictive pdf based on Viterbi approximation (Jiang et al., 1997, 1999),

$$\tilde{p}(\boldsymbol{X}|\varphi, W) \approx \max_{\boldsymbol{s},\boldsymbol{l}} \int f(\boldsymbol{X}, \boldsymbol{s}, \boldsymbol{l}|\Lambda, W) \cdot p(\Lambda|\varphi, W)\,\mathrm{d}\Lambda. \tag{3}$$

The resultant BPC decision rule is called VBPC rule,

$$\hat{W} = \arg\max_{W} \left[p(W) \cdot \max_{\boldsymbol{s},\boldsymbol{l}} \int f(\boldsymbol{X}, \boldsymbol{s}, \boldsymbol{l}|\Lambda, W) \cdot p(\Lambda|\varphi, W)\,\mathrm{d}\Lambda \right]. \tag{4}$$

A detailed recursive search algorithm to implement Eq. (4) can be found in (Jiang et al., 1997, 1999).

From our previous study, we notice that the prior pdf $p(\Lambda|\varphi, W)$ plays a key role in the VBPC rule. If $p(\Lambda|\varphi, W)$ adequately describes the mismatches between the training and testing conditions, the robustness of recognition systems can be greatly improved even when the mismatches in question are quite large. In

(Jiang et al., 1997, 1999), we have examined a so-called *less-informative* (actually constrained uniform) prior distribution to deal with some unknown mismatches. The less-informative prior pdf has a simple functional form, but its hyperparameters can not be easily estimated in advance and the prior pdf itself is also difficult to be updated when new knowledge becomes available. In order to incorporate new information dynamically into the existing system, in this section, we first extend the VBPC formulation to another class of prior pdf which belongs to the family of natural conjugate prior pdf's (see (Gauvain and Lee, 1994) for the details) of the complete-data density of CDHMM.

Assuming that we model each word W with an N-state CDHMM with a parameter vector $\Lambda = (\pi, A, \theta)$, where π is the initial state distribution, A is the transition matrix, and θ is the parameter vector composed of mixture parameters $\theta_i = \{\omega_{ik}, m_{ik}, r_{ik}\}_{k=1,2,\ldots,K}$ for each state i (K denotes the number of mixtures in each state), with the mixture coefficients ω_{ik}, the D-dimensional mean vectors m_{ik}, and the $D \times D$ precision (inverse covariance) matrices r_{ik}. In this paper, we only consider the uncertainty of the mean vectors in CDHMM with diagonal precision matrices. Therefore, the natural conjugate prior pdf of the complete-data density has a Gaussian functional form,

$$p(\Lambda|\varphi, W) = \prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{d=1}^{D}\sqrt{\frac{\tau_{ikd}}{2\pi}}\exp\left[-\frac{1}{2}\tau_{ikd}(m_{ikd}-\mu_{ikd})^2\right], \tag{5}$$

where $\varphi = \{\mu_{ikd}, \tau_{ikd} | 1 \leqslant i \leqslant N,\ 1 \leqslant k \leqslant K,\ 1 \leqslant d \leqslant D\}$ are hyperparameters.

Substitute the prior pdf Eq. (5) into Eq. (3), under the condition that only the mean vectors of CDHMM are considered to be uncertain, the Viterbi approximate predictive density can be simply computed as (see Appendix A for more details)

$$\tilde{p}(\boldsymbol{X}|\varphi, W) \approx f(\boldsymbol{X}, \boldsymbol{s}^*, \boldsymbol{l}^*|\Lambda, W) \cdot \frac{p(\Lambda|\varphi, W)}{p(\Lambda|\varphi', W)}, \tag{6}$$

where $\varphi' = \{\mu'_{ikd}, \tau'_{ikd} | 1 \leqslant i \leqslant N,\ 1 \leqslant k \leqslant K,\ 1 \leqslant d \leqslant D\}$ are the updated hyperparameters which will be explained in Eqs. (11) and (12), and $\{\boldsymbol{s}^*, \boldsymbol{l}^*\}$ denote the optimal path found via VBPC search algorithm (shown later in Eq. (9)).

One implementation issue here is the hyperparameter estimation of the initial prior pdf's, i.e., how to design suitable prior pdf's from the available parameters of the pre-trained CDHMM's. Following the idea in (Huo and Lee, 1997a–c), we use the initialization method as follows:

$$\mu_{ikd}^{(0)} = m_{ikd} \text{ and } \tau_{ikd}^{(0)} = \epsilon \cdot r_{ikd} \cdot c_{ik} \cdot g_d, \tag{7}$$

where $\epsilon > 0$ is a weighting coefficient, c_{ik} is a weight count accumulated for the kth mixture component of the state i during training CDHMM's parameters. $g_d = d^2 \cdot \rho^d\,(\rho > 1.0)$. According to the upper bound of a perturbation in speech cepstral domain given in (Merhav and Lee, 1993), $g_d = d^2 \cdot \rho^d\,(\rho > 1.0)$ is used to avoid too severe smoothing in the high dimension of the cepstral vector.

3. Sequential segmental Bayesian learning for VBPC

Since the performance of VBPC based on the initial prior pdf's constructed as Eq. (7) usually is not good enough, we follow Huo and Lee (1997b,c) to adopt Bayesian learning to update the prior pdf's in order to improve VBPC's performance successively.

Given the initial prior pdf's $p(\Lambda|W)$ and independent observation samples $\boldsymbol{X}^{(n)} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$, the formal sequential Bayesian learning is performed as follows (Huo and Lee, 1997a):

$$p\big(\Lambda|\boldsymbol{X}^{(n)}, W\big) = \frac{f(\boldsymbol{X}_n|\Lambda, W) \cdot p\big(\Lambda|\boldsymbol{X}^{(n-1)}, W\big)}{\int_\Omega f(\boldsymbol{X}_n|\Lambda, W) \cdot p\big(\Lambda|\boldsymbol{X}^{(n-1)}, W\big)\,\mathrm{d}\Lambda}. \tag{8}$$

where Ω denotes an admissible region of the parameter space. Starting the calculation from $p(\Lambda|\boldsymbol{X}^{(0)}, W) = p(\Lambda|W)$, we can obtain a sequence of prior/posterior densities $p(\Lambda|\boldsymbol{X}^{(1)}, W)$, $p(\Lambda|\boldsymbol{X}^{(2)}, W)$, and so forth, with gradually increased accuracy. Once these updated pdf's are obtained, they are employed in VBPC procedure in place of the initial prior pdf's $p(\Lambda|W)$.

However, there is no closed form solution to the above sequential learning procedure for CDHMM (see discussions in (Huo and Lee, 1997a)). In practice, some approximations are needed. In this paper, we first study a simple sequential Bayesian learning strategy, namely segmental Bayesian learning, for VBPC. Then, we extend segmental Bayesian learning to cover a finite mixture approximation of the CDHMM's true prior/posterior density, by which we attempt to derive a series of prior pdf's with increased accuracy to describe the unknown mismatches in robust speech recognition.

In (Huo and Lee, 1997a), under a very general framework, a specific approximation procedure, namely quasi-Bayes (QB) learning method has been proposed and extensively studied. Starting from the Viterbi version of the QB procedure in (Huo and Lee, 1997a), in this paper, we try to get the *optimal* state and mixture component label sequences by using the labeling algorithm embedded in VBPC approach (Jiang et al., 1997, 1999) as follows:

$$\{\boldsymbol{s}^*, \boldsymbol{l}^*\} = \arg\max_{\boldsymbol{s},\boldsymbol{l}} \int_{\Omega} f(\boldsymbol{X}, \boldsymbol{s}, \boldsymbol{l}|\Lambda, W) \cdot p(\Lambda|W)\,\mathrm{d}\Lambda. \tag{9}$$

In this way, we are trying to find the hidden label sequences to maximize the joint predictive density of the observation and hidden label sequences instead of the conventional joint density as in standard Viterbi labeling algorithm. Once the intended hidden label sequences are identified, we can use the same formulations as in (Huo and Lee, 1997a) to update the posterior pdf $p(\Lambda|\boldsymbol{X}, W)$.

Given an adaptation data $\boldsymbol{X} = \{X_1, X_2, \ldots, X_T\}$, we first determine the optimal pair $\{\boldsymbol{s}^*, \boldsymbol{l}^*\}$ as in Eq. (9). Then the posterior pdf $p(\Lambda|\boldsymbol{X}, W)$ is approximated as follows according to segmental (or Viterbi) Bayesian learning:

$$p(\Lambda|\boldsymbol{X}, W) \propto f(\boldsymbol{X}, \boldsymbol{s}^*, \boldsymbol{l}^*|\Lambda, W) \cdot p(\Lambda|W). \tag{10}$$

If the natural conjugate prior pdf $p(\Lambda|W)$ like Eq. (5) is chosen, the posterior pdf also has the form as the right hand side of Eq. (5), with the adapted hyperparameters μ'_{ikd} and τ'_{ikd} given as follows:

$$\mu'_{ikd} = \frac{\mu_{ikd}\tau_{ikd} + \mu^*_{ikd}\tau^*_{ikd}}{\tau_{ikd} + \tau^*_{ikd}}, \tag{11}$$

$$\tau'_{ikd} = \tau_{ikd} + \tau^*_{ikd}, \tag{12}$$

where

$$\mu^*_{ikd} = \frac{\sum_{t=1}^{T} X_{td}\delta(\boldsymbol{s}^*_t - i)\delta(\boldsymbol{l}^*_t - k)}{\sum_{t=1}^{T} \delta(\boldsymbol{s}^*_t - i)\delta(\boldsymbol{l}^*_t - k)}, \tag{13}$$

$$\tau^*_{ikd} = r_{ikd}\sum_{t=1}^{T}\delta(\boldsymbol{s}^*_t - i)\delta(\boldsymbol{l}^*_t - k) \tag{14}$$

and $\delta(\cdot)$ is the Kronecher delta function.

4. Sequential Bayesian learning of CDHMM based on a finite mixture approximation of its prior/posterior pdf

The above segmental Bayesian learning is easy to perform, and the updated prior/posterior pdf's always keep the simple form and unimodal shape. However, the unimodal functional form might be too simple for the prior pdf to describe the unknown mismatches usually encountered in robust speech recognition.

A more flexible prior/posterior pdf might be useful to make BPC approach perform better. A natural choice is to use a finite mixture distribution for the prior/posterior pdf (e.g., Smith and Makov, 1985; Titterington et al., 1985; Bernardo and Giron, 1988). In this section, we extend the above segmental Bayesian learning to cover a finite mixture prior/posterior pdf. In the next section, an N-Best based implementation is derived to incrementally adapt the finite mixture prior/posterior pdf to new data so that the performance of VBPC can be successively improved.

According to Eq. (8), the true posterior pdf after observing $\boldsymbol{X}$ can be expressed as

$$p(\Lambda|\boldsymbol{X},W) \propto f(\boldsymbol{X}|\Lambda,W)\cdot p(\Lambda|W) = \sum_{\iota\in\Upsilon} f(\boldsymbol{X},\iota|\Lambda,W)\cdot p(\Lambda|W), \tag{15}$$

where for convenience ι, called a *path*, denotes a combination pair of $\{\boldsymbol{s},\boldsymbol{l}\}$, and the path space Υ consists of all possible ι.

We further examine the predictive density of $\boldsymbol{X}$,

$$\begin{aligned}\tilde{p}(\boldsymbol{X}|\varphi,W) &= \int f(\boldsymbol{X}|\Lambda,W)\cdot p(\Lambda|\varphi,W)\,\mathrm{d}\Lambda \\ &= \sum_{\iota\in\Upsilon}\int f(\boldsymbol{X},\iota|\Lambda,W)\cdot p(\Lambda|W)\,\mathrm{d}\Lambda = \sum_{\iota\in\Upsilon}\varpi(\boldsymbol{X}|\iota,W),\end{aligned} \tag{16}$$

where

$$\varpi(\boldsymbol{X}|\iota,W) = \int f(\boldsymbol{X},\iota|\Lambda,W)\cdot p(\Lambda|W)\,\mathrm{d}\Lambda. \tag{17}$$

Here $\varpi(\boldsymbol{X}|\iota,W)$ denotes the component part of the predictive density corresponding to the path ι in Υ, which can be computed via VBPC algorithm in (Jiang et al., 1997, 1999). We notice that the true posteriori pdf in Eq. (15) is a finite mixture function, which consists of numerous homogeneous terms. Each term in turn corresponds to a path in Υ. It is reasonable to pick up the M most significant terms among Υ, based on their contribution to the predictive density, i.e. $\varpi(\boldsymbol{X}|\iota,W)$, to approximate the true posterior pdf and truncate others in order to keep computation and memory under control. That is,

$$\Xi^{(M)} = \arg\max_{\iota\in\Upsilon}^{(M)} \varpi(\boldsymbol{X}|\iota,W), \tag{18}$$

where $\arg\max^{(M)}$ denotes the operation to choose the M largest items, $\Xi^{(M)}$ denotes the set of the M most significant terms. Then the approximate posterior pdf can be expressed as

$$p(\Lambda|\boldsymbol{X},W) \approx \frac{\sum_{\iota\in\Xi^{(M)}} f(\boldsymbol{X},\iota|\Lambda,W)\cdot p(\Lambda|W)}{\sum_{\iota\in\Xi^{(M)}}\varpi(\boldsymbol{X}|\iota,W)} = \sum_{\iota\in\Xi^{(M)}} \varepsilon_\iota \cdot p(\Lambda|\iota,\boldsymbol{X},W), \tag{19}$$

where

$$\varepsilon_\iota = \frac{\varpi(\boldsymbol{X}|\iota,W)}{\sum_{\iota\in\Xi^{(M)}}\varpi(\boldsymbol{X}|\iota,W)} \tag{20}$$

and $p(\Lambda|\iota,\boldsymbol{X},W)$ denotes natural conjugate prior of the complete-data density given ι, whose form has been shown in Eq. (5).

5. N-Best based implementation

Assuming that we have observed training data $\boldsymbol{X}^{(n-1)}$, the current prior/posterior pdf follows Eq. (19) and can be shown as

$$p(\Lambda|\boldsymbol{X}^{(n-1)},W) = \sum_{\iota_1\in\Xi_1^{(M)}} \varepsilon_{\iota_1}\cdot p(\Lambda|\boldsymbol{X}^{(n-1)},\iota_1,W)$$
$$= \sum_{\iota_1\in\Xi_1^{(M)}} \varepsilon_{\iota_1}\cdot \prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{d=1}^{D}\sqrt{\frac{\tau_{ikd}^{(\iota_1)}}{2\pi}}\exp\left[-\frac{1}{2}\tau_{ikd}^{(\iota_1)}\left(m_{ikd}-\mu_{ikd}^{(\iota_1)}\right)^2\right], \tag{21}$$

where $\tau_{ikd}^{(\iota_1)}$ and $\mu_{ikd}^{(\iota_1)}$ are hyperparameters.

When a new data $\boldsymbol{X}_n=\{X_{n1},X_{n2},\ldots,X_{nT}\}$ becomes available, we first use N-Best VBPC algorithm to decode the top M best paths. We denote the set of these M best paths as $\Xi_2^{(M)}$. According to Eq. (9), the prior $p(\Lambda|\varphi,W)$ determines the VBPC search. Thus the top M best paths found by N-Best VBPC search algorithm are different from the M best paths decoded from the normal N-Best Viterbi search. Then, for each path in $\Xi_2^{(M)}$, we can derive the corresponding component of the likelihood function. Thus the current likelihood function can be approximated as a summation of M mixtures, i.e.,

$$f(\boldsymbol{X}_n|\Lambda,W)\approx\sum_{\iota_2\in\Xi_2^{(M)}} f(\boldsymbol{X}_n,\iota_2|\Lambda,W)=\sum_{\iota_2\in\Xi_2^{(M)}} C^{(\iota_2)}\cdot\prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{d=1}^{D}\exp\left[-\frac{1}{2}\tau_{ikd}^{(\iota_2)}\left(m_{ikd}-\mu_{ikd}^{(\iota_2)}\right)^2\right], \tag{22}$$

with

$$\mu_{ikd}^{(\iota_2)}=\frac{\sum_{t=1}^{T}X_{ntd}\delta\left(\boldsymbol{s}_t^{(\iota_2)}-i\right)\delta\left(\boldsymbol{l}_t^{(\iota_2)}-k\right)}{\sum_{t=1}^{T}\delta\left(\boldsymbol{s}_t^{(\iota_2)}-i\right)\delta\left(\boldsymbol{l}_t^{(\iota_2)}-k\right)}, \tag{23}$$

$$\tau_{ikd}^{(\iota_2)}=r_{ikd}\sum_{t=1}^{T}\delta\left(\boldsymbol{s}_t^{(\iota_2)}-i\right)\delta\left(\boldsymbol{l}_t^{(\iota_2)}-k\right), \tag{24}$$

$$C^{(\iota_2)}=\pi_{\boldsymbol{s}_1^{(\iota_2)}}\omega_{\boldsymbol{s}_1^{(\iota_2)}\boldsymbol{l}_1^{(\iota_2)}}\sqrt{\frac{r_{\boldsymbol{s}_1^{(\iota_2)}\boldsymbol{l}_1^{(\iota_2)}}}{2\pi}}\prod_{t=2}^{T}a_{\boldsymbol{s}_{t-1}^{(\iota_2)}\boldsymbol{s}_t^{(\iota_2)}}\omega_{\boldsymbol{s}_t^{(\iota_2)}\boldsymbol{l}_t^{(\iota_2)}}\sqrt{\frac{r_{\boldsymbol{s}_t^{(\iota_2)}\boldsymbol{l}_t^{(\iota_2)}}}{2\pi}}$$
$$\times\prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{d=1}^{D}\exp\left[-\frac{r_{ikd}}{2}\sum_{t=1}^{T}\left[\left(X_{ntd}^2-\left(\mu_{ikd}^{(\iota_2)}\right)^2\right)\delta\left(\boldsymbol{s}_t^{(\iota_2)}-i\right)\delta\left(\boldsymbol{l}_t^{(\iota_2)}-k\right)\right]\right], \tag{25}$$

where $\boldsymbol{s}_t^{(\iota)}$ and $\boldsymbol{l}_t^{(\iota)}$ denotes the state and Gaussian component labels corresponding to time instant t in the path ι, respectively.

According to Bayes' theorem in Eq. (8), the new posterior pdf corresponds to the product of the prior pdf and the likelihood. Based on our approximations of prior pdf and likelihood function in Eqs. (21) and (22) respectively, the new posterior pdf $p(\Lambda|\boldsymbol{X}^{(n)},W)$ includes M^2 terms, which is denoted here as the set $\Xi^{(M^2)}$. Each term of $\Xi^{(M^2)}$ corresponds to a product of each ι_1 in $\Xi_1^{(M)}$ and each ι_2 in $\Xi_2^{(M)}$. We denote it as ι, i.e. $\iota=\iota_1\otimes\iota_2$. Then

$$p(\Lambda|\boldsymbol{X}^{(n)},W)\propto\sum_{\iota\in\Xi^{(M^2)}}\varpi\left(\boldsymbol{X}_n|\boldsymbol{X}^{(n-1)},\iota,W\right)\cdot p\left(\Lambda|\boldsymbol{X}^{(n)},\iota,W\right), \tag{26}$$

where

$$\varpi\left(\boldsymbol{X}_n|\boldsymbol{X}^{(n-1)},\iota,W\right)=\varepsilon_{\iota_1}\times C^{(\iota_2)}\times\prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{d=1}^{D}\sqrt{\frac{\tau_{ikd}^{(\iota_1)}}{\tau_{ikd}^{(\iota_1)}+\tau_{ikd}^{(\iota_2)}}}\cdot\exp\left[-\frac{1}{2}\frac{\tau_{ikd}^{(\iota_1)}\tau_{ikd}^{(\iota_2)}}{\tau_{ikd}^{(\iota_1)}+\tau_{ikd}^{(\iota_2)}}\left(\mu_{ikd}^{(\iota_1)}-\mu_{ikd}^{(\iota_2)}\right)^2\right] \tag{27}$$

and $p(\Lambda|\boldsymbol{X}^{(n)}, \iota, W)$ has the same form as $p(\Lambda|\boldsymbol{X}^{(n-1)}, \iota_1, W)$ in Eq. (21), with the adapted hyperparameters $\tau_{ikd}^{(\iota)}$ and $\mu_{ikd}^{(\iota)}$ given as follows:

$$\tau_{ikd}^{(\iota)} = \tau_{ikd}^{(\iota_1)} + \tau_{ikd}^{(\iota_2)}, \tag{28}$$

$$\mu_{ikd}^{(\iota)} = \frac{\mu_{ikd}^{(\iota_1)} \cdot \tau_{ikd}^{(\iota_1)} + \mu_{ikd}^{(\iota_2)} \cdot \tau_{ikd}^{(\iota_2)}}{\tau_{ikd}^{(\iota_1)} + \tau_{ikd}^{(\iota_2)}}. \tag{29}$$

In order to reduce the computational and storage overhead, we still choose the M most significant terms from $\Xi^{(M^2)}$ based on $\varpi(\boldsymbol{X}_n|\boldsymbol{X}^{(n-1)}, \iota, W)$, i.e.

$$\Xi^{(M)} = \arg\max_{\iota \in \Xi^{(M^2)}}^{(M)} \varpi\big(\boldsymbol{X}_n|\boldsymbol{X}^{(n-1)}, \iota, W\big). \tag{30}$$

Thus we approximate the posterior distribution $p(\Lambda|\boldsymbol{X}^{(n)}, W)$ by these M terms,

$$p\big(\Lambda|\boldsymbol{X}^{(n)}, W\big) \approx \frac{\sum_{\iota \in \Xi^{(M)}} \varpi\big(\boldsymbol{X}_n|\boldsymbol{X}^{(n-1)}, \iota, W\big) \cdot p\big(\Lambda|\boldsymbol{X}^{(n)}, \iota, W\big)}{\sum_{\iota \in \Xi^{(M)}} \varpi\big(\boldsymbol{X}_n|\boldsymbol{X}^{(n-1)}, \iota, W\big)} = \sum_{\iota \in \Xi^{(M)}} \varepsilon_\iota \cdot p\big(\Lambda|\boldsymbol{X}^{(n)}, \iota, W\big), \tag{31}$$

where

$$\varepsilon_\iota = \frac{\varpi\big(\boldsymbol{X}_n|\boldsymbol{X}^{(n-1)}, \iota, W\big)}{\sum_{\iota \in \Xi^{(M)}} \varpi\big(\boldsymbol{X}_n|\boldsymbol{X}^{(n-1)}, \iota, W\big)}. \tag{32}$$

The updated posterior pdf $p(\Lambda|\boldsymbol{X}^{(n)}, W)$ can be used in Eq. (4) in place of $p(\Lambda|\varphi, W)$ to improve VBPC's performance. As a remark, we can see that the above sequential Bayesian learning strategy is a generalization of segmental Bayesian learning, which is included as a special case ($M = 1$). The N-Best implementation flow of the above sequential Bayesian learning algorithm is sketched in Fig. 2.

As a remark, the N-Best method has been used in adaptation by Matsui and Furui (1998). However, in our work, N-Best strategy is used to derive a finite mixture form of the prior/posterior pdf. And the derived mixture prior/posterior pdf's are evolved under the theoretical Bayesian framework. Thus the work in this paper are significantly different from that of Matsui and Furui (1998). Besides, in (Mokbel, 1997), an N-Best solution is also used to consider alternatives to the optimal path in unsupervised mode and is combined with stochastic matching for equalization.

One important issue in N-Best implementation is related to the choice of top M mixands in the finite mixture approximation. In practice, if the chosen mixands are too similar to each other (it is the case especially when the mixands are derived from N-Best paths as in the above N-Best implementation), the finite mixture approximation of the posterior pdf can not provide more information than a unimodal approximation. A heuristic solution to mitigate the problem is to merge those similar mixands during the N-Best approximation process as described below. Let the mixands $f(\boldsymbol{X}_n, \iota_2|\Lambda, W)$ in Eq. (22) be indexed by $\iota_2^{(1)}, \iota_2^{(2)}, \ldots, \iota_2^{(M)}$, which correspond to the top M most significant mixands in $\Xi_2^{(M)}$ in order. The dissimilarity measure, $d(\iota_2^{(m)}, \iota_2^{(n)})$, between two mixands is simply defined and computed by directly checking the path difference between two paths of $\iota_2^{(m)}$ and $\iota_2^{(n)}$.

IF $d(\iota_2^{(m)}, \iota_2^{(n)}) \leqslant \vartheta_1$, where we assume $m < n$ and ϑ_1 is a preset threshold;
THEN we merge mixand $\iota_2^{(n)}$ with $\iota_2^{(m)}$:
(i) to remove mixand $\iota_2^{(n)}$,
(ii) to update the weight of $\iota_2^{(m)}$ as $C^{\iota_2^{(m)}} = \vartheta_2 \cdot (C^{\iota_2^{(m)}} + C^{\iota_2^{(n)}})$, where $\vartheta_2 > 0$ is another preset constant to control the merging.

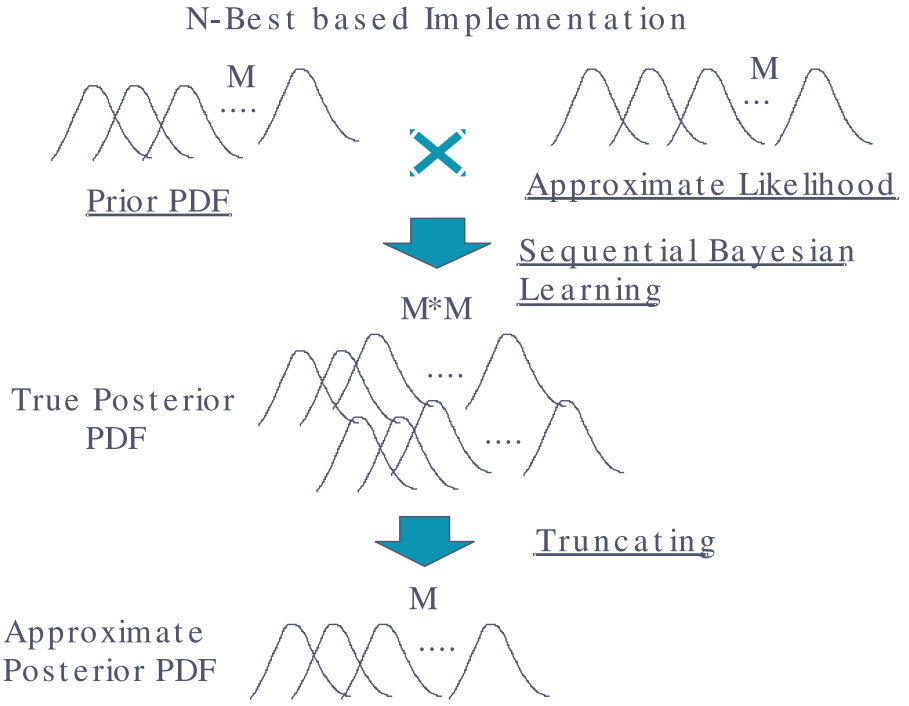

Fig. 2. *N*-Best implementation of the sequential Bayesian learning based on the finite mixture approximation of the true prior/posterior pdf.

By choosing the control parameters ϑ_1 and ϑ_2 appropriately, we can obtain the needed mixture approximation of the posterior pdf.

6. Experimental results

To examine the viability of the above algorithm, it is applied to a speaker-independent (SI) recognition task of isolated Japanese digits where the unknown mismatch exists between training and testing conditions. We have studied two types of mismatch: (i) the mismatch caused by additive white Gaussian noise, (ii) cross-gender mismatch. The speech data are selected from ATR Japanese Speech Database. It contains 0–9 Japanese digit utterances from 60 speakers (half male, half female). Each digit is modeled by a left-to-right 4-state CDHMM without state skipping and each state has six Gaussian mixture components with diagonal covariance matrices. Each feature vector consists of 16 LPC-derived cepstral coefficients. In the following experiments, two control parameters ϑ_1 and ϑ_2 are manually set in advance, and remain constant during the adaptation procedure because we can not find an easy way to adjust them automatically.

6.1. Noisy speech recognition

One mismatch to be examined is caused by additive noise. While SI training is performed on clean speech data, computer-generated Gaussian white noise is added to the testing and adaptation data with the same level of intensity prior to the preprocessing. The experimental results are shown in Fig. 3. In Fig. 3,

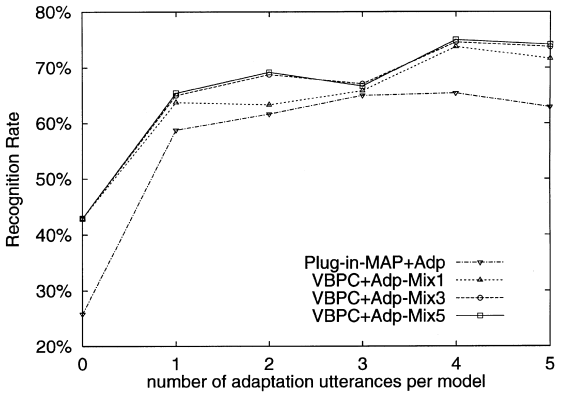

Fig. 3. Performance comparison of noisy speech recognition at SNR = 20 dB as a function of amount of adaptation data among methods by combining sequential Bayesian learning with plug-in MAP decoding and VBPC (with mixture number $M = 1, 3, 5$).

"Plug-in-MAP+Adp" denotes that a plug-in MAP decision rule is used in speech recognition and an on-line Bayesian learning algorithm is used to adapt CDHMM's parameters; "VBPC+Adp-Mix1", "VBPC+Adp-Mix3" and "VBPC+Adp-Mix5" denote that the VBPC decision rule is used in speech recognition and the prior/posterior pdf of the CDHMM's is approximated by one (i.e., segmental Bayesian learning), three and five mixture pdf's, respectively, in each step of on-line adaptation. It is shown that VBPC method surpasses the conventional plug-in MAP decision rule when no knowledge about mismatch is available at the beginning, where the prior pdf's are initialized as in Eq. (7). The performance of VBPC can be further improved via incremental adaptation of the prior/posterior pdf with the new adaptation data. It is observed that VBPC consistently outperforms the plug-in MAP decoding in this case. In addition, a better performance of VBPC can be achieved by using three mixture components in the prior/posterior pdf than a unimodal pdf if the pdf mixands are appropriately pruned and merged as described above. But only a slight improvement has been observed when we further increase mixture number from three to five.

6.2. Cross-gender speech recognition

We have also examined a more general mismatch caused by gender difference. In the cross-gender experiments, we train the CDHMMs with all the female speech data. The male speech data are divided into two sets. One is used for adaptation and another for testing. The experimental results are shown in Fig. 4. A similar learning behavior is observed here as that in noisy speech recognition. We observe that the initial improvement of VBPC over plug-in MAP rule without any adaptation data is minor comparing to that in noisy speech recognition. However, a significant improvement has been observed when we replace unimodal pdf with a three-mixture pdf. It suggests that mixture approximation helps more when dealing with a more complex mismatch situation.

6.3. Convergence property of the sequential bayesian learning

The convergence property of the sequential Bayesian learning in terms of the recognition accuracy improvement based on VBPC and plug-in MAP decoding in noisy speech recognition is displayed in Fig. 5. The results show that the on-line Bayesian learning schemes maintain a good asymptotic convergence property in both VBPC and plug-in MAP decision rules.

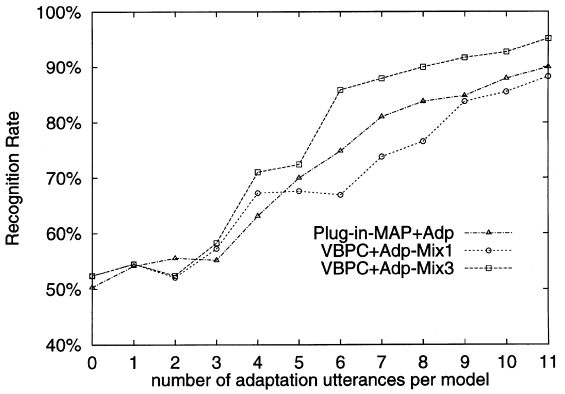

Fig. 4. Performance comparison of cross-gender speech recognition as a function of amount of adaptation data among methods by combining sequential Bayesian learning with plug-in MAP decoding and VBPC (with mixture number $M = 1, 3$).

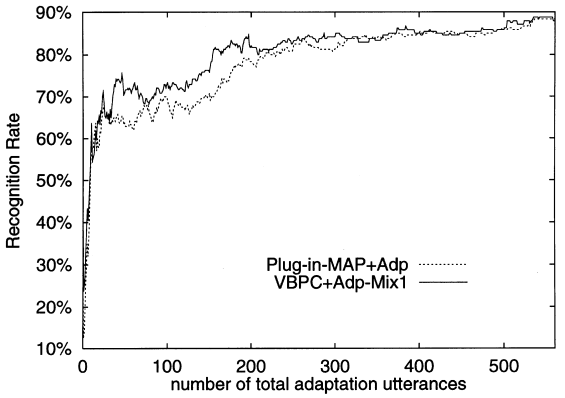

Fig. 5. Convergence property comparison at SNR $= 20$ dB among methods by combining sequential Bayesian learning with plug-in MAP decoding and VBPC (with mixture number $M = 1$).

7. Discussion and conclusion

Theoretically speaking, whether VBPC achieves a satisfactory performance or not greatly depends on whether the prior pdf can really reflect the mismatch in question. Ideally, such a prior pdf should be constructed from the subject knowledge about the possible mismatch involved in real applications. However, in practice, it seems more attractive if we can use an automatic learning method to elicit the required prior/posterior pdf from some training data. In this study, we show again that the natural conjugate prior pdf of the complete-data density of CDHMM is a good choice for the initial prior pdf. Although it might not be good enough initially, we show that the prior pdf can be dynamically improved via sequential Bayesian learning. The experimental results confirm that it is helpful to use a finite mixture approximation in both Bayesian learning and VBPC calculation. Such an improvement over the unimodal pdf approximation greatly depends on how properly the true pdf is pruned. Concretely, how to automatically adjust some control parameters, namely ϑ_1 and ϑ_2, during the adaptation procedure. This is an important issue which is still under our investigation. Furthermore, in the current implementation, the new

precision information incorporated in Bayesian learning procedure, namely $\tau_{ikd}^{(t_2)}$ in Eq. (40), is directly derived from pre-trained model's precision as in Eq. (24). Thus we can not warrant that the updated posterior pdf's reflect the mismatch more accurately. To take uncertainty of both mean and precision parameters simultaneously into account might be helpful. Moreover, the Bayesian learning algorithm proposed in this paper is suitable and useful to update the system to track the slow condition changes. Obviously the strategy is not applicable to those abrupt changes. We need other mechanisms to deal with these problems, which is beyond the scope of this paper. As a final remark, the sequential learning of a mixture distribution, which has no sufficient statistics with a fixed dimension, is a quite challenging problem. Although the formal Bayesian learning theoretically converges to the optimal solution under the condition of unlimited memory and calculation, only suboptimal methods can be implemented in practice. The N-Best implementation studied here is sensitive to the pruning, selection, and merging of the mixture components in the sequential adaptation procedure. More efforts are still needed to look for a better prior pdf in BPC approach.

Appendix A

By adopting Viterbi approximation, the predictive pdf can be expressed as in Eq. (3),

$$\begin{aligned}\tilde{p}(\boldsymbol{X}|\varphi, W) &\approx \max_{\boldsymbol{s},\boldsymbol{l}} \int f(\boldsymbol{X}, \boldsymbol{s}, \boldsymbol{l}|\Lambda, W) \cdot p(\Lambda|\varphi, W)\, \mathrm{d}\Lambda \\ &= \int f(\boldsymbol{X}, \boldsymbol{s}^*, \boldsymbol{l}^*|\Lambda, W) \cdot p(\Lambda|\varphi, W)\, \mathrm{d}\Lambda, \end{aligned} \tag{33}$$

where

$$\{\boldsymbol{s}^*, \boldsymbol{l}^*\} = \arg\max_{\boldsymbol{s},\boldsymbol{l}} \int f(\boldsymbol{X}, \boldsymbol{s}, \boldsymbol{l}|\Lambda, W) \cdot p(\Lambda|\varphi, W)\, \mathrm{d}\Lambda \tag{34}$$

can be achieved by using VBPC search algorithm in (Jiang et al., 1997, 1999).

In case we only consider the uncertainty of the mean vectors in CDHMM with diagonal precision matrices, the natural conjugate prior pdf of the complete-data density has a Gaussian functional form as in Eq. (5) and is repeated here,

$$p(\Lambda|\varphi, W) = \prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{d=1}^{D}\sqrt{\frac{\tau_{ikd}}{2\pi}}\exp\left[-\frac{1}{2}\tau_{ikd}(m_{ikd} - \mu_{ikd})^2\right], \tag{35}$$

where $\varphi = \{\mu_{ikd},\ \tau_{ikd} | 1 \leqslant i \leqslant N,\ 1 \leqslant k \leqslant K,\ 1 \leqslant d \leqslant D\}$ are hyperparameters.

On the other hand, the likelihood function corresponding to the optimal path $\{\boldsymbol{s}^*, \boldsymbol{l}^*\}$ can be expressed as

$$f(\boldsymbol{X}, \boldsymbol{s}^*, \boldsymbol{l}^*|\Lambda, W) = C^* \cdot \prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{d=1}^{D}\exp\left[-\frac{1}{2}\tau_{ikd}^*\left(m_{ikd} - \mu_{ikd}^*\right)^2\right], \tag{36}$$

with

$$\mu_{ikd}^* = \frac{\sum_{t=1}^{T} X_{td}\delta(\boldsymbol{s}_t^* - i)\delta\left(\boldsymbol{l}_t^* - k\right)}{\sum_{t=1}^{T}\delta\left(\boldsymbol{s}_t^* - i\right)\delta\left(\boldsymbol{l}_t^* - k\right)}, \tag{37}$$

$$\tau_{ikd}^* = r_{ikd}\sum_{t=1}^{T}\delta\left(\boldsymbol{s}_t^* - i\right)\delta\left(\boldsymbol{l}_t^* - k\right). \tag{38}$$

According to Bayes' theorem, the approximate posterior pdf $p(\Lambda|\varphi', W)$ is the product of the approximate likelihood function $f(\boldsymbol{X}, \boldsymbol{s}^*, \boldsymbol{l}^*|\Lambda, W)$ and the prior distribution $p(\Lambda|\varphi, W)$. If the the normalization factor is taken into account, we have

$$p(\Lambda|\varphi', W) = \prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{d=1}^{D}\sqrt{\frac{\tau'_{ikd}}{2\pi}}\exp\left[-\frac{1}{2}\tau'_{ikd}\left(m_{ikd}-\mu'_{ikd}\right)^2\right], \tag{39}$$

with

$$\tau'_{ikd} = \tau_{ikd} + \tau^*_{ikd}, \tag{40}$$

$$\mu'_{ikd} = \frac{\mu_{ikd}\cdot\tau_{ikd} + \mu^*_{ikd}\cdot\tau^*_{ikd}}{\tau_{ikd}+\tau^*_{ikd}}. \tag{41}$$

Substitute Eqs. (35) and (36) into Eq. (33), we have

$$\begin{aligned}
\tilde{p}(\boldsymbol{X}|\varphi, W) &\approx C^*\prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{d=1}^{D}\int\sqrt{\frac{\tau_{ikd}}{2\pi}}\exp\left[-\frac{1}{2}\tau_{ikd}(m_{ikd}-\mu_{ikd})^2\right]\cdot\exp\left[-\frac{1}{2}\tau^*_{ikd}\left(m_{ikd}-\mu^*_{ikd}\right)^2\right]\mathrm{d}m_{ikd}\\
&= C^*\prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{d=1}^{D}\sqrt{\frac{\tau_{ikd}}{2\pi}}\cdot\exp\left[-\frac{1}{2}\frac{\tau_{ikd}\cdot\tau^*_{ikd}}{\tau_{ikd}+\tau^*_{ikd}}\left(\mu_{ikd}-\mu^*_{ikd}\right)^2\right]\cdot\int\exp\left[-\frac{1}{2}\tau'_{ikd}\left(m_{ikd}-\mu'_{ikd}\right)^2\right]\mathrm{d}m_{ikd}\\
&= C^*\prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{d=1}^{D}\sqrt{\frac{\tau_{ikd}}{\tau'_{ikd}}}\cdot\exp\left[-\frac{1}{2}\frac{\tau_{ikd}\cdot\tau^*_{ikd}}{\tau_{ikd}+\tau^*_{ikd}}\left(\mu_{ikd}-\mu^*_{ikd}\right)^2\right]\\
&= f(\boldsymbol{X}, \boldsymbol{s}^*, \boldsymbol{l}^*|\Lambda, W)\cdot\frac{p(\Lambda|\varphi, W)}{p(\Lambda|\varphi', W)}.
\end{aligned} \tag{42}$$

References

Bernardo, J.M., Giron, F.J., 1988. A Bayesian analysis of simple mixture problems. In: Bernardo, J.M., DeGroot, M.H., Lindley, D.V., Smith, A.F.M. (Eds.), Bayesian Statistics 3. Oxford University Press, London, pp. 67–78.

Furui, S., 1997. Recent advances in robust speech recognition. In: Proceedings of the ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels. Pont-a-Mousson, France, pp. 11–20.

Gauvain, J.-L., Lee, C.-H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Transactions on Speech and Audio Processing 2 (2), 291–298.

Huo, Q., Lee, C.-H., 1997a. On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate. IEEE Transactions on Speech and Audio Processing 5 (2), 161–172.

Huo, Q., Lee, C.-H., 1997b. Combined on-line model adaptation and Bayesian predictive classification for robust speech recognition. In: Proceedings of European Conference on Speech Communication and Technology, Rhodes, Greece, pp. 1847–1850.

Huo, Q., Lee, C.-H., 1997c. A Bayesian predictive classification approach to robust speech recognition. IEEE Transactions on Speech and Audio Processing, submitted.

Huo, Q., Jiang, H., Lee, C.-H., 1997. A Bayesian predictive classification approach to robust speech recognition. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing, Munich, Germany, pp. II-1547–1550.

Jiang, H., Hirose, K., Huo, Q., 1997. Robust speech recognition based on Viterbi Bayesian predictive classification. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing, Munich, Germany, pp. II-1551–1554.

Jiang, H., Hirose, K., Huo, Q., 1999. Robust speech recognition based on Bayesian prediction approach. IEEE Transactions on Speech and Audio Processing 7 (4) (in press).

Lee, C.-H., 1998. On stochastic feature and model compensation approaches to robust speech recognition. Speech Communication 25 (1–3), 29–47.

Matsui, T., Furui, S., 1998. N-Best-based unsupervised speaker adaptation for speech recognition. Computer Speech and Language 12 (1), 41–50.

Merhav, N., Lee, C.-H., 1993. A minimax classification approach with application to robust speech recognition. IEEE Transactions on Speech and Audio Processing 1 (1), 90–100.

Mokbel, C., 1997. MUSE: MUlti-path Stochastic Equalization: A theoretical framework to combine equalization and stochastic modeling. In: Proceedings of the ESCA Workshop on Robust Speech Recognition, Pont-a-Mousson, France, p. 211.

Smith, A.F.M., Makov, U.E., 1985. Bayesian detection and estimation of jumps in linear systems. In: Bernardo, J.M. et al. (Eds.), Bayesian Statistics 2, Elsevier, Amsterdam.

Titterington, D.M., Smith, A.F.M., Makov, U.E., 1985. Statistical Analysis of Finite Mixture Distributions. Wiley, London.