IMPROVING VITERBI BAYESIAN PREDICTIVE CLASSIFICATION VIA SEQUENTIAL BAYESIAN LEARNING IN ROBUST SPEECH RECOGNITION

Hui Jiang[†] *Keikichi Hirose*[†] *Qiang Huo*[‡]

[†]Department of Information and Communication Engineering, University of Tokyo, Japan
[‡] Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong

ABSTRACT

In this paper, we extend our previously proposed Viterbi Bayesian predictive classification (VBPC) algorithm to accommodate a new class of prior probability density function (pdf) for continuous density hidden Markov model (CDHMM) based robust speech recognition. The initial prior pdf of CDHMM is assumed to be a finite mixture of natural conjugate prior pdf's of its complete-data density. With the new observation data, the true posterior pdf is approximated by the same type of finite mixture pdf's which retain the required most significant terms in the true posterior density according to their contribution to the corresponding predictive density. Then the updated mixture pdf is used to improve the VBPC performance. The experimental results on a speaker-independent recognition task of isolated Japanese digits confirm the viability and the usefulness of the proposed technique.

1. INTRODUCTION

In order to deal with the possible modeling/estimation errors and/or unknown mismatches between training and testing conditions, we have been investigating a *Bayesian predictive classification* (BPC) approach for robust speech recognition [2, 3, 4, 5]. In this approach, we use a quite general prior pdf (*probability density function*) $p(\Lambda|\varphi)$ to characterize the variability of the model parameter Λ caused by the abovementioned distortions. We try to average out this variability while making decision for speech recognition and such a BPC rule operates as follows:

$$\hat{W} = \arg\max_{W} \tilde{p}(W|\mathbf{X}) = \arg\max_{W} \tilde{p}(\mathbf{X}|W) \cdot P_{\Gamma}(W) \quad (1)$$

where $\mathbf{X}$ is the observed feature vector sequence to be recognized, $P_{\Gamma}(W)$ is the language model with parameter Γ,

$$\tilde{p}(\mathbf{X}|W) = \int_{\Omega} f(\mathbf{X}|\Lambda, W) p(\Lambda|\varphi, W) \mathrm{d}\Lambda \quad (2)$$

is called the *predictive pdf* of the observation $\mathbf{X}$ given the symbol sequence W, $f(\mathbf{X}|\Lambda, W)$ is the conventional acoustic model with parameters Λ, and $\hat{W}$ is the recognized symbol (usually word) sequence of interest embedded in the observation sequence $\mathbf{X}$. For a Gaussian mixture continuous density hidden Markov model (CDHMM) based speech recognition system, we have to use some approximation methods to compute the predictive density in Eq. (2) and thus propose a Viterbi approximation method in [5] as follows:

$$\tilde{p}(\mathbf{X}|W) \simeq \max_{s,l} \int f(\mathbf{X}, s, l|\Lambda, W) p(\Lambda|\varphi, W) \mathrm{d}\Lambda \quad (3)$$

where s is the unobserved state sequence and l is the associated sequence of the unobserved mixture component labels corresponding to the observation sequence $\mathbf{X}$. A detailed recursive search algorithm to implement Eq.(3) can be found in [5]. We observed that an appropriate prior pdf is crucial for BPC based robust speech recognition. In [5], a constrained uniform prior distribution was adopted. In spite of its simple functional form, it is difficult to estimate its initial hyperparameters and to update them even when new data/knowledge become available. As already shown in [2, 3, 4], by adopting a family of natural conjugate prior pdf's of the *complete-data* density of CDHMM, BPC helps in many types of distortions [2]. If we can access some adaptation data, by combining BPC with the data-driven on-line Bayesian adaptation techniques [1], the prior pdf can be made more appropriate and thus the robustness of the speech recognition system can be further enhanced [3]. Furthermore, the *knowledge* and/or *experience* of the interaction between speech signal and the possible mismatch can also be used to guide us to obtain a better prior pdf which can improve the BPC performance as shown in [4].

Motivated by the works in [1, 2, 3, 4], in this paper, we extend our VBPC formulation to accommodate not only the abovementioned conjugate prior pdf's of the *complete-data* density of CDHMM, but also their *finite mixtures*. It is this finite mixture approximation approach that this paper focuses on. More specifically, the initial prior pdf of CDHMM is assumed to be a finite mixture of natural conjugate prior pdf's of its *complete-data* density. With the new observation data, the true posterior pdf is approximated by the same type of finite mixture pdf's. In the operation, we use an N-best search algorithm to retain the required most significant terms in the true posterior density according to their contribution to the corresponding predictive density. In this way, a more accurate prior/posterior pdf of the HMM parameters can be obtained and hopefully the VBPC performance is improved. The above Bayesian model adaptation and VBPC decoding strategy is applied to a speaker-independent recognition task of isolated Japanese digits to deal with two types of mismatch between training and testing conditions: i) the mismatch caused by additive white

0-7803-4428-6/98 $10.00 © 1998 IEEE

Gaussian noise, ii) cross-gender mismatch. In the following sections, we will describe the details of the the proposed method and the experimental results, and finally conclude with a confirmation of the viability and the usefulness of the proposed techniques.

2. SEQUENTIAL BAYESIAN LEARNING OF CDHMM BASED ON FINITE MIXTURE APPROXIMATION OF POSTERIOR PDF

For the simplicity of the discussion, we consider the isolated word recognition where each word is modeled by an N-state CDHMM with 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 $\theta_i = \{\omega_{ik}, m_{ik}, r_{ik}\}_{k=1,2,\cdots,K}$ for each state i, with the mixture coefficients ω_{ik}, the mean vectors m_{ik} , and the precision (inverse covariance) matrices r_{ik}. Given initial prior pdf $p(\Lambda|W)$ and observation samples $\mathbf{X}^n = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, the formal sequential Bayesian learning is performed as follows:

$$p(\Lambda|\mathbf{X}^n, W) = \frac{f(\mathbf{x}_n|\Lambda, W) \cdot p(\Lambda|\mathbf{X}^{n-1}, W)}{\int_\Omega f(\mathbf{x}_n|\Lambda, W) \cdot p(\Lambda|\mathbf{X}^{n-1}, W) d\Lambda} \quad (4)$$

where Ω denotes an admissible region of the parameter space, and $f(\mathbf{x}_n|\Lambda, W)$ is the likelihood function. Starting the calculation from $p(\Lambda|\mathbf{X}^0, W) = p(\Lambda|W)$, we can obtain a sequence of prior/posterior densities $p(\Lambda|\mathbf{X}^1, W)$, $p(\Lambda|\mathbf{X}^2, W)$, and so forth, with gradually increased accuracy [1]. Theoretically speaking, the *posterior* pdf after observing $\mathbf{x}$ can be computed as

$$p(\Lambda|\mathbf{x}, W) \propto p(\Lambda|W) \cdot f(\mathbf{x}|\Lambda, W) = \sum_{\iota \epsilon \Upsilon} p(\Lambda|W) \cdot f(\mathbf{x}, \iota|\Lambda, W) \quad (5)$$

where ι, called a *path*, denotes a combination of a state path s and a mixture component label sequence l, and the path space Υ consists of all possible ι. We further examine the predictive density of $\mathbf{x}$

$$f(\mathbf{x}|W) = \int p(\Lambda|W) \cdot f(\mathbf{x}|\Lambda, W)\, d\Lambda = \sum_{\iota \epsilon \Upsilon} \int p(\Lambda|W) \cdot f(\mathbf{x}, \iota\,|\Lambda, W)\, d\Lambda = \sum_{\iota \epsilon \Upsilon} \varpi(\mathbf{x}|\iota, W) \quad (6)$$

where $\varpi(\mathbf{x}|\iota, W) = \int p(\Lambda|W) \cdot f(\mathbf{x}, \iota\,|\Lambda, W)\, d\Lambda$. $\varpi(\mathbf{x}|\iota, W)$ denotes the component part of the predictive density corresponding to the *path* ι in Υ, which can be computed via VBPC algorithm in [5]. We notice that the true *posteriori* pdf in Eq. (5) 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. $\varpi(\mathbf{x}|\iota, W)$, to approximate the true posterior pdf and truncate others in order to keep computation and memory under control. That is, $\Xi^{(M)} = \arg\max^{(M)}_{\iota \epsilon \Upsilon} \varpi(\mathbf{x}|\iota)$, 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

$$p(\Lambda|\mathbf{x}, W) \approx \frac{\sum_{\iota \epsilon \Xi^{(M)}} f(\mathbf{x}, \iota|\Lambda, W) \cdot p(\Lambda|W)}{\sum_{\iota \epsilon \Xi^{(M)}} \varpi(\mathbf{x}|\iota, W)} = \sum_{\iota \epsilon \Xi^{(M)}} \omega_\iota \cdot p(\Lambda|\iota, \mathbf{x}, W) \quad (7)$$

where $\omega_\iota = \frac{\varpi(\mathbf{x}|\iota, W)}{\sum_{\iota \epsilon \Xi^{(M)}} \varpi(\mathbf{x}|\iota, W)}$, and $p(\Lambda|\iota, \mathbf{x}, W)$ denotes natural conjugate prior of the complete-data density given ι, whose form will be explained later.

3. N-BEST BASED IMPLEMENTATION

As a first step, we only consider the uncertainty of the mean vectors in CDHMM. Assuming that we have observed training data $\mathbf{X}^{(n-1)}$, the current prior/posterior pdf follows Eq.(7) and can be shown as

$$p(\Lambda|\mathbf{X}^{(n-1)}, W) = \sum_{\iota_1 \epsilon \Xi_1^{(M)}} \omega_{\iota_1} \cdot p(\Lambda|\mathbf{X}^{(n-1)}, \iota_1, W) = \sum_{\iota_1 \epsilon \Xi_1^{(M)}} \omega_{\iota_1} \cdot \prod_{i=1}^{N} \prod_{k=1}^{K} \prod_{d=1}^{D} \sqrt{\frac{\tau_{ikd}^{(\iota_1)}}{2\pi}} e^{-\frac{1}{2}\tau_{ikd}^{(\iota_1)}(m_{ikd} - \mu_{ikd}^{(\iota_1)})^2} \quad (8)$$

where $\tau_{ikd}^{(\iota_1)}$ and $\mu_{ikd}^{(\iota_1)}$ are hyperparameters. The above equation also gives the form of natural conjugate prior pdf of the complete-data density given ι_1 when only mean vectors of CDHMM are random. When a new data $\mathbf{x}_n$ becomes available, the current likelihood function can be approximately calculated by N-best VBPC algorithm and also expressed as a summation of M mixtures, i.e.

$$f(\mathbf{x}_n|\Lambda, W) \approx \sum_{\iota_2 \epsilon \Xi_2^{(M)}} f(\mathbf{x}_n, \iota_2|\Lambda, W) = \sum_{\iota_2 \epsilon \Xi_2^{(M)}} C^{(\iota_2)} \cdot \prod_{i=1}^{N} \prod_{k=1}^{K} \prod_{d=1}^{D} e^{-\frac{1}{2}\tau_{ikd}^{(\iota_2)}(m_{ikd} - \mu_{ikd}^{(\iota_2)})^2} \quad (9)$$

where

$$\mu_{ikd}^{(\iota_2)} = \frac{\sum_{t=1}^{T} x_{ntd} \delta(s_t^{(\iota_2)} - i)\delta(l_t^{(\iota_2)} - k)}{\sum_{t=1}^{T} \delta(s_t^{(\iota_2)} - i)\delta(l_t^{(\iota_2)} - k)} \quad (10)$$

$$\tau_{ikd}^{(\iota_2)} = r_{ikd} \sum_{t=1}^{T} \delta(s_t^{(\iota_2)} - i)\delta\,(l_t^{(\iota_2)} - k) \quad (11)$$

$$C^{(\iota_2)} = \pi_{s_1^{(\iota_2)}\omega_{s_1^{(\iota_2)}l_1^{(\iota_2)}}} \sqrt{\frac{\tau_{s_1^{(\iota_2)}l_1^{(\iota_2)}}}{2\pi}} \prod_{t=2}^{T} a_{s_{t-1}^{(\iota_2)}s_t^{(\iota_2)}\omega_{s_t^{(\iota_2)}\,l_t^{(\iota_2)}}} \sqrt{\frac{\tau_{s_t^{(\iota_2)}l_t^{(\iota_2)}}}{2\pi}} \prod_{i=1}^{N} \prod_{k=1}^{K} \prod_{d=1}^{D} \exp[-\frac{r_{ikd}}{2} \sum_{t=1}^{T} [(x_{ntd}^2 - \mu_{ikd}^{(\iota_2)^2})\delta(s_t^{(\iota_2)} - i)\delta(l_t^{(\iota_2)} - k)]] \quad (12)$$

From Eq.(4), the new *posterior* pdf $p(\Lambda|\mathbf{X}^n, W)$ includes M^2 terms, denoted here as the set $\Xi^{(M^2)}$. Each term of $\Xi^{(M^2)}$ corresponds to a combination 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

$$p(\Lambda|\mathbf{X}^n, W) \propto \sum_{\iota \in \Xi^{(M^2)}} \varpi(\mathbf{x}_n|\mathbf{X}^{(n-1)}, \iota, W) \cdot p(\Lambda|\mathbf{X}^n, \iota, W) \tag{13}$$

where

$$\varpi(\mathbf{x}_n|\mathbf{X}^{(n-1)}, \iota, W) = w_{\iota_1} \times C^{(\iota_2)} \times \prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{d=1}^{D} \sqrt{\frac{\tau_{ikd}^{(\iota_1)}}{\tau_{ikd}^{(\iota_1)} + \tau_{ikd}^{(\iota_2)}}} \cdot \exp[-\frac{1}{2}\frac{\tau_{ikd}^{(\iota_1)}\tau_{ikd}^{(\iota_2)}}{\tau_{ikd}^{(\iota_1)} + \tau_{ikd}^{(\iota_2)}}(\mu_{ikd}^{(\iota_1)} - \mu_{ikd}^{(\iota_2)})^2] \tag{14}$$

and $p(\Lambda|\mathbf{X}^n, \iota, W)$ has the same form as $p(\Lambda|\mathbf{X}^{(n-1)}, \iota_1, W)$ in Eq.(8), with the adapted hyperparameters $\tau_{ikd}^{(\iota)}$ and $\mu_{ikd}^{(\iota)}$ given as follows:

$$\tau_{ikd}^{(\iota)} = \tau_{ikd}^{(\iota_1)} + \tau_{ikd}^{(\iota_2)} \tag{15}$$

$$\mu_{ikd}^{(\iota)} = \frac{\mu_{ikd}^{(\iota_1)} \cdot \tau_{ikd}^{(\iota_1)} + \mu_{ikd}^{(\iota_2)} \cdot \tau_{ikd}^{(\iota_2)}}{\tau_{ikd}^{(\iota_1)} + \tau_{ikd}^{(\iota_2)}} \tag{16}$$

In order to reduce the computational and storage overhead, we still choose the M most significant terms from $\Xi^{(M^2)}$ based on $\varpi(\mathbf{x}_n|\mathbf{X}^{(n-1)}, \iota, W)$, i.e. $\Xi^{(M)} = \arg\max_{\iota \in \Xi^{(M^2)}}^{(M)} \varpi(\mathbf{x}_n|\mathbf{X}^{(n-1)}, \iota, W)$, and approximate the *posterior* distribution $p(\Lambda|\mathbf{X}^n, W)$ by these M terms:

$$p(\Lambda|\mathbf{X}^n, W) \approx \frac{\sum_{\iota \in \Xi^{(M)}} \varpi(\mathbf{x}_n|\mathbf{X}^{(n-1)}, \iota, W) \cdot p(\Lambda|\mathbf{X}^n, \iota, W)}{\sum_{\iota \in \Xi^{(M)}} \varpi(\mathbf{x}_n|\mathbf{X}^{(n-1)}, \iota, W)} = \sum_{\iota \in \Xi^{(M)}} w_\iota \cdot p(\Lambda|\mathbf{X}^n, \iota, W) \tag{17}$$

where $w_\iota = \frac{\varpi(\mathbf{x}_n|\mathbf{X}^{(n-1)}, \iota, W)}{\sum_{\iota \in \Xi^{(M)}} \varpi(\mathbf{x}_n|\mathbf{X}^{(n-1)}, \iota, W)}$. The updated posterior pdf $p(\Lambda|\mathbf{X}^n, W)$ can be used in Eq.(3) in place of $p(\Lambda|\varphi, W)$ to improve VBPC's performance.

4. IMPLEMENTATION ISSUES

One implementation issue is the hyperparameter estimation of the initial prior pdf, i.e., how to design a suitable prior pdf from available parameters of the pre-trained CDHMM's. Following the ideas in [1, 2, 3, 4], we use the initialization method as follows: $\mu_{ikd}^{(0)} = m_{ikd}$ and $\tau_{ikd}^{(0)} = \epsilon \cdot r_{ikd} \cdot c_{ik} \cdot g_d$,where $\epsilon > 0$ is a weighting coefficient, c_{ik} is a weight count accumulated for the k-th mixture component of the state i during training CDHMM's parameters, and $g_d = d^2 \cdot \rho^d$ $(\rho > 1.0)$ is used to avoid over smoothing in higher dimension of the mean vector.

Another issue is related to the choice of top N 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 pathes 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 $f(\mathbf{x}_n, \iota_2|\Lambda, W)$ in Eq.(9) be indexed by $\iota_2^{(1)}, \iota_2^{(2)}, \cdots, \iota_2^{(M)}$, 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 pathes of $\iota_2^{(m)}$ and $\iota_2^{(n)}$.

IF $d(\iota_2^{(m)}, \iota_2^{(n)}) \leq \varepsilon_1$, where we assume $m < n$ and ε_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 $C^{\iota_2^{(m)}} = \varepsilon_2 \cdot (C^{\iota_2^{(m)}} + C^{\iota_2^{(n)}})$, where $\varepsilon_2 > 0$ is another preset constant to control the merging.

By choosing the control parameters ε_1 and ε_2 appropriately, we can obtain the needed mixture approximation of the posterior pdf.

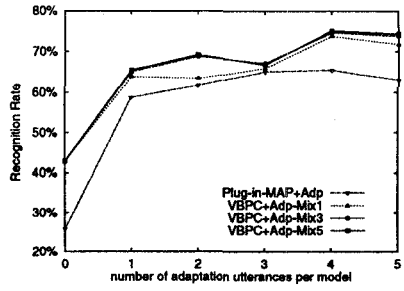

Figure 1. Performance comparison of noisy speech recognition at SNR = 20(dB) as a function of amount of adaptation data among methods in which sequential Bayesian learning is combined with plug-in MAP decoding or VBPC (with mixture number $M = 1, 3, 5$)

5. EXPERIMENTAL RESULTS

To examine the viability of the above algorithm, it was 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 is 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 6 Gaussian mixture components with diagonal covariance matrices. Each feature vector consists of 16 LPC-derived cepstral coefficients.

5.1. 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 Figure 1, where "Plug-in-MAP+Adp" denotes that we use plug-in MAP decision rule

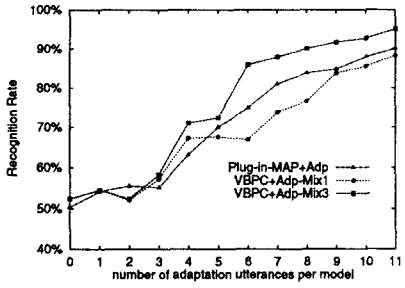

Figure 2. Performance comparison of cross-gender speech recognition as a function of amount of adaptation data among methods in which sequential Bayesian learning is combined with plug-in MAP decoding or VBPC (with mixture number $M = 1, 3$)

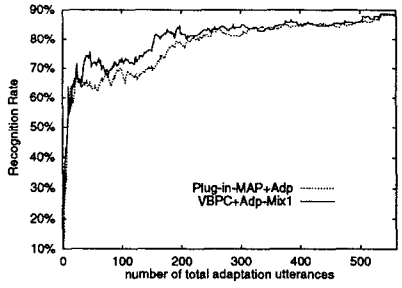

Figure 3. Convergence property comparison at $SNR = 20(dB)$ among methods in which sequential Bayesian learning is combined with plug-in MAP decoding or VBPC (with mixture number $M = 1$)

in speech recognition and an on-line Bayesian learning algorithm (see [1] for details) to adapt CDHMMs' parameters, and where "VBPC+Adp-Mix1", "VBPC+Adp-Mix3", and "VBPC+Adp-Mix5" denote that VBPC decision rule is used in speech recognition and the prior/posterior pdf of CDHMM is approximated by one, three, and five mixture pdf's respectively in each step of 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. The performance of VBPC can be further improved via incremental adaptation of the prior/posterior pdf continuously with 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.

5.2. 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 Figure 2. A similar learning behavior is observed here as the one 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 larger improvement has been observed when we replace unimodal pdf with three-mixture pdf. It suggests that mixture approximation helps more when dealing with a more complex mismatch situation.

5.3. Convergence Property of VBPC

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 Figure 3. The results show that the on-line Bayesian learning schemes maintain a good asymptomatic convergence property in both VBPC and plug-in MAP decision rules.

6. DISCUSSION AND CONCLUSION

The experimental results show that it is helpful to use a finite mixture approximation in both Bayesian learning and VBPC calculation. The improvement greatly depends on how properly the true pdf is pruned. Furthermore, in the current implementation, the new precision information incorporated in Bayesian learning procedure, namely $\tau_{ikd}^{(\iota_2)}$ in Eq.(15), is directly derived from pre-trained model's precision as in Eq.(11). 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. The sequential learning of a mixture distribution, which has no sufficient statistics with a fixed dimension, seems to be 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 mixands pruning, selection, and merging in the sequential adaptation procedure. More efforts are still needed to look for a better prior pdf in BPC approach.

REFERENCES

[1] Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate," *IEEE Trans. on SAP*, Vol. 5, pp.161-172, 1997.

[2] Q. Huo, H. Jiang and C.-H. Lee, "A Bayesian predictive classification approach to robust speech recognition," *Proc. ICASSP-97*, pp.II-1547-1550.

[3] Q. Huo and C.-H. Lee, "Combined on-line model adaptation and Bayesian predictive classification for robust speech recognition," *Proc. EUROSPEECH-97*, pp.1847-1850.

[4] Q. Huo and C.-H. Lee, "A study of prior sensitivity for Bayesian predictive classification based robust speech recognition," submitted to *ICASSP-98*, October 1997.

[5] H. Jiang, K. Hirose and Q. Huo, "Robust speech recognition based on Viterbi Bayesian predictive classification," in *Proc. ICASSP-97*, pp.II-1551-1554, 1997.