SEQUENTIAL BAYESIAN LEARNING OF CDHMM BASED ON FINITE MIXTURE APPROXIMATION OF ITS PRIOR/POSTERIOR DENSITY

Hui Jiang[†], Keikichi Hirose[†] and Qiang Huo[‡]
[†]Department of Information and Communication Engineering, University of Tokyo, Japan
[‡]ATR Interpreting Telecommunications Research Labs., 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

Abstract - In this paper, we propose a sequential Bayesian learning strategy of CDHMM based on finite mixture approximation of its prior/posterior density. The initial prior density of CDHMM is assumed to be a finite mixture of natural conjugate prior pdf's of the 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 Bayesian predictive density by using an N-best beam search algorithm. Then the updated mixture pdf is used in VBPC method to deal with unknown mismatches in robust speech recognition. The experimental results on a speaker-independent recognition task of isolated Japanese digits confirm the viability and the usefulness of the proposed method.

1. INTRODUCTION

In order to deal with unknown mismatches between training and testing conditions, we have investigated a *Bayesian predictive classification* (BPC) approach in [2, 3, 4, 5] for robust speech recognition. We observed that an appropriate prior probability density function (pdf) is crucial for BPC based robust speech recognition. Motivated by the works in [1, 2, 3, 4], in this paper, we aim at improving the BPC performance by adopting and sequentially adapting a more accurate prior/posterior distribution of the HMM parameters in a Gaussian mixture continuous density HMM (CDHMM) based speech recognition system. The initial prior density of CDHMM is assumed to be a finite mixture of natural conjugate prior pdf's of the *complete-data* density.

0-7803-3698-4/97/$10.00 © 1997 IEEE.

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 Bayesian predictive density by using an N-best beam search algorithm. The above Bayesian adaptation strategy has been 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 Gaussian noise, ii) cross-gender mismatch. The experimental results confirm the viability and the usefulness of the proposed method.

2. Sequential Bayesian Learning of CDHMM

We model each speech unit (referred to as *word* heretofore) with 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}. Assume our initial knowledge about CDHMM parameters Λ of word W is contained in *a priori* pdf $p(\Lambda|W)$. Given independent observation samples $\mathbf{X}^n = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, the formal sequential Bayesian learning is performed as follows:

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

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. However, there is no closed form solution to the above sequential learning procedure for CDHMM. In practice, some approximations are needed. In this paper, we study a sequential Bayesian learning strategy for CDHMM based on finite mixture approximation of its prior/posterior density.

3. Finite Mixture Approximation of Posterior PDF

Let's examine the likelihood function of CDHMM Λ of word W after observing an data $\mathbf{x}$,

$$f(\mathbf{x}|\Lambda, W) = \sum_{s,l} f(\mathbf{x}, s, l\,|\Lambda, W) = \sum_{\iota \epsilon \Upsilon} f(\mathbf{x}, \iota\,|\Lambda, W) \tag{2}$$

where the summations are taken over all possible state path s and mixture component label sequence l. For convenience, we name a combination of s and l as a *path* ι. The path space Υ consists of all possible ι. Therefore, the *posterior* pdf after observing $\mathbf{x}$ can be computed by eq.(1) as

$$p(\Lambda|\mathbf{x}, W) \propto p(\Lambda|W) \cdot f(\mathbf{x}|\Lambda, W) = \sum_{\iota \epsilon \Upsilon} p(\Lambda|W) \cdot f(\mathbf{x}, \iota|\Lambda, W) \quad (3)$$

We further examine the Bayesian predictive density of $\mathbf{x}$

$$\begin{aligned} f(\mathbf{x}|W) &= \int p(\Lambda|W) \cdot f(\mathbf{x}|\Lambda, W)\, \mathrm{d}\Lambda \\ &= \sum_{\iota \epsilon \Upsilon} \int p(\Lambda|W) \cdot f(\mathbf{x}, \iota|\Lambda, W)\, \mathrm{d}\Lambda = \sum_{\iota \epsilon \Upsilon} \varpi(\mathbf{x}|\iota, W) \end{aligned} \quad (4)$$

where $\varpi(\mathbf{x}|\iota, W) = \int p(\Lambda|W) \cdot f(\mathbf{x}, \iota|\Lambda, W)\, \mathrm{d}\Lambda$. $\varpi(\mathbf{x}|\iota, W)$ denotes the component part of predictive density corresponding to the *path* ι in Υ, which can be easily computed via Viterbi BPC (VBPC) algorithm in [5].

We notice that the true *posteriori* pdf (3) 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,

$$\Xi^{(M)} = \underset{\iota \epsilon \Upsilon}{\operatorname{argmax}}^{(M)} \varpi(\mathbf{x}|\iota) \quad (5)$$

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

$$p(\Lambda|\mathbf{x}, W) \approx \frac{\sum_{\iota \epsilon \Xi^{(M)}} f(\mathbf{x}, \iota|\Lambda, W) \cdot p(\Lambda|W)}{\sum_{\iota \epsilon \Xi^{(M)}} \varpi(\mathbf{x}|\iota, W)} = \sum_{\iota \epsilon \Xi^{(M)}} \omega_\iota \cdot p(\Lambda|\iota, \mathbf{x}, W) \quad (6)$$

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.

4. 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.(6) and can be shown as

$$\begin{aligned} p(\Lambda|\mathbf{X}^{(n-1)}, W) &= \sum_{\iota_1 \epsilon \Xi_1^{(M)}} \omega_{\iota_1} \cdot p(\Lambda|\mathbf{X}^{(n-1)}, \iota_1, W) \\ &= \sum_{\iota_1 \epsilon \Xi_1^{(M)}} \omega_{\iota_1} \cdot \prod_{i=1}^{N} \prod_{k=1}^{K} \prod_{d=1}^{D} \sqrt{\frac{\tau_{ikd}^{(\iota_1)}}{2\pi}} e^{-\frac{1}{2}\tau_{ikd}^{(\iota_1)}(m_{ikd} - \mu_{ikd}^{(\iota_1)})^2} \end{aligned} \quad (7)$$

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.

$$f(\mathbf{x}_n|\Lambda, W) \approx \sum_{\iota_2 \epsilon \Xi_2^{(M)}} f(\mathbf{x}_n, \iota_2|\Lambda, W)$$

$$= \sum_{\iota_2 \epsilon \Xi_2^{(M)}} C^{(\iota_2)} \cdot \prod_{i=1}^{N} \prod_{k=1}^{K} \prod_{d=1}^{D} e^{-\frac{1}{2}\tau_{ikd}^{(\iota_2)}(m_{ikd}-\mu_{ikd}^{(\iota_2)})^2} \tag{8}$$

where

$$\mu_{ikd}^{(\iota_2)} = \frac{\sum_{t=1}^{T} x_{ntd}\delta(s_t^{(\iota_2)} - i)\delta(l_t^{(\iota_2)} - k)}{\sum_{t=1}^{T} \delta(s_t^{(\iota_2)} - i)\delta(l_t^{(\iota_2)} - k)} \tag{9}$$

$$\tau_{ikd}^{(\iota_2)} = r_{ikd} \sum_{t=1}^{T} \delta(s_t^{(\iota_2)} - i)\delta\ (l_t^{(\iota_2)} - k) \tag{10}$$

$$C^{(\iota_2)} = \pi_{s_1^{(\iota_2)}\omega_{s_1^{(\iota_2)}l_1^{(\iota_2)}}} \sqrt{\frac{r_{s_1^{(\iota_2)}l_1^{(\iota_2)}}}{2\pi}} \prod_{t=2}^{T} a_{s_{t-1}^{(\iota_2)}s_t^{(\iota_2)}} \omega_{s_t^{(\iota_2)}\ l_t^{(\iota_2)}} \sqrt{\frac{r_{s_t^{(\iota_2)}l_t^{(\iota_2)}}}{2\pi}}$$

$$\prod_{i=1}^{N} \prod_{k=1}^{K} \prod_{d=1}^{D} \exp[-\frac{r_{ikd}}{2} \sum_{t=1}^{T} [(x_{ntd}^2 - {\mu_{ikd}^{(\iota_2)}}^2)\delta(s_t^{(\iota_2)} - i)\delta(l_t^{(\iota_2)} - k)]] \tag{11}$$

According to eq.(1), the new *posterior* pdf $p(\Lambda|\mathbf{X}^n, W)$ includes M^2 terms (denoted as the set $\Xi^{(M^2)}$), each of which 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. $\iota = \iota_1 \otimes \iota_2$, and

$$p(\Lambda|\mathbf{X}^n, W) \propto \sum_{\iota \epsilon \Xi^{(M^2)}} \varpi(\mathbf{x}_n|\mathbf{X}^{(n-1)}, \iota, W) \cdot p(\Lambda|\mathbf{X}^n, \iota\ , W) \tag{12}$$

where

$$\varpi(\mathbf{x}_n|\mathbf{X}^{(n-1)}, \iota, W) = w_{\iota_1} \times C^{(\iota_2)}$$

$$\times \prod_{i=1}^{N} \prod_{k=1}^{K} \prod_{d=1}^{D} \sqrt{\frac{\tau_{ikd}^{(\iota_1)}}{\tau_{ikd}^{(\iota)}}} \cdot \exp[-\frac{1}{2} \frac{\tau_{ikd}^{(\iota_1)}\tau_{ikd}^{(\iota_2)}}{\tau_{ikd}^{(\iota_1)} + \tau_{ikd}^{(\iota_2)}} (\mu_{ikd}^{(\iota_1)} - \mu_{ikd}^{(\iota_2)})^2] \tag{13}$$

and $p(\Lambda|\mathbf{X}^n, \iota, W)$ has the same form as $p(\Lambda|\mathbf{X}^{(n-1)}, \iota_1, W)$ in eq.(7), with the adapted hyperparameters $\tau_{ikd}^{(\iota)}$ and $\mu_{ikd}^{(\iota)}$ given as follows:

$$\tau_{ikd}^{(\iota)} = \tau_{ikd}^{(\iota_1)} + \tau_{ikd}^{(\iota_2)} \tag{14}$$

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

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:

$$\begin{aligned} p(\Lambda|\mathbf{X}^n, W) &\approx \frac{\sum_{\iota \in \Xi^{(M)}} \varpi(\mathbf{x}_n|\mathbf{X}^{(n-1)}, \iota, W) \cdot p(\Lambda|\mathbf{X}^n, \iota, W)}{\sum_{\iota \in \Xi^{(M)}} \varpi(\mathbf{x}_n|\mathbf{X}^{(n-1)}, \iota, W)} \\ &= \sum_{\iota \in \Xi^{(M)}} w_\iota \cdot p(\Lambda|\mathbf{X}^n, \iota, W) \end{aligned} \tag{16}$$

where $w_\iota = \frac{\varpi(\mathbf{x}_n|\mathbf{X}^{(n-1)}, \iota, W)}{\sum_{\iota \in \Xi^{(M)}} \varpi(\mathbf{x}_n|\mathbf{X}^{(n-1)}, \iota, W)}$.

5. Viterbi Bayesian Predictive Classification (VBPC)

After updating the posterior pdf $p(\Lambda|\mathbf{X}^n, W)$, we can recognize any new incoming data, denoted as $\mathbf{y}$, via VBPC approach as follows:

$$\hat{W} = \underset{W}{\text{argmax}} \max_{s,l} \int f(\mathbf{y}, s, l|\Lambda, W) \cdot p(\Lambda|\mathbf{X}^n, W)\, \mathrm{d}\Lambda \tag{17}$$

A detailed recursive search algorithm to implement eq.(17) can be found in [5].

6. Implementation Issues

One 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 before we observe any new data. Like in [2, 3], we use the initialization method proposed in [1] as follows: $\mu_{ikd}^{(0)} = m_{ikd}$, $\tau_{ikd}^{(0)} = \epsilon \cdot r_{ikd} \cdot c_{ik}$, where $\epsilon > 0$ is a weighting coefficient, and c_{ik} is a weight count accumulated for the k-th mixture component of the state i during training of CDHMM's parameters.

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.(8) be indexed by $\iota_2^{(1)}, \iota_2^{(2)}, \cdots, \iota_2^{(M)}$, which correspond to the 1st, 2nd, $\cdots$, M-th most significant mixands in $\Xi_2^{(M)}$ respectively. 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

$$d(\iota_2^{(m)}, \iota_2^{(n)}) \leq \varepsilon_1, \tag{18}$$

where we assume $m < n$ and ε_1 is a preset threshold;

THEN we merge mixand $\iota_2^{(n)}$ with $\iota_2^{(m)}$, i.e. to remove the mixand $\iota_2^{(n)}$ and update weight of $\iota_2^{(m)}$ as

$$C^{\iota_2^{(m)}} = \varepsilon_2 \cdot (C^{\iota_2^{(m)}} + C^{\iota_2^{(n)}}) \tag{19}$$

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.

7. Experiments and Discussions

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). The speech was recorded in a quiet environment at sampling rate of 20kHz with 16bit quantization. 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.

7.1. Noisy Speech Recognition

A simple special case of mismatch situation is encountered when the testing signal is corrupted by various additive noises, while the training data are clean. 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 in speech recognition and an on-line Bayesian learning algorithm (see [1] for details) to adapt CDHMM-s' 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 density respectively in each step of adaptation. It is shown that the performance of VBPC can be improved via incremental adaptation of the prior/posterior pdf with new data. It is also observed that VBPC consistently outperforms the conventional plug-in MAP decoding in this case. Given the same amount of adaptation data, a better performance of VBPC

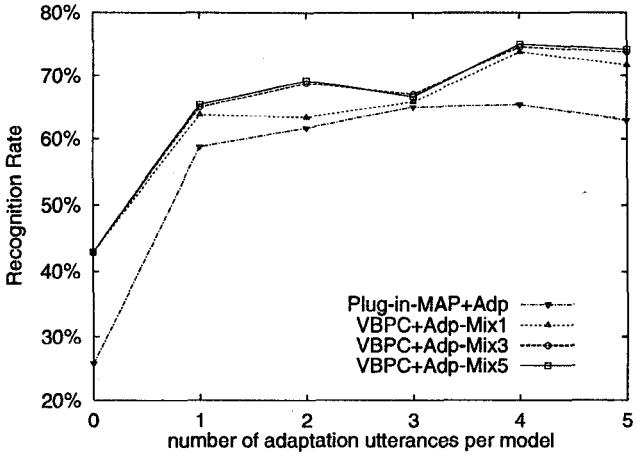

Figure 1: Performance comparison of noisy speech recognition at SNR = 20(dB) as a function of amount of adaptation data among methods in which sequential Bayesian learning is combined with plug-in MAP decoding or VBPC (with mixture number $M = 1, 3, 5$)

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 in every adaptation step. Only a slight improvement has been observed when we further increase mixture number from three to five.

7.2 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. But in this case, a bigger 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.

8. Final Remarks

The experimental results show that it is helpful to use a finite mixture approximation in both Bayesian learning and BPC calculation. The improvement greatly depends on how properly the true pdf is pruned. The sequential estimation 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

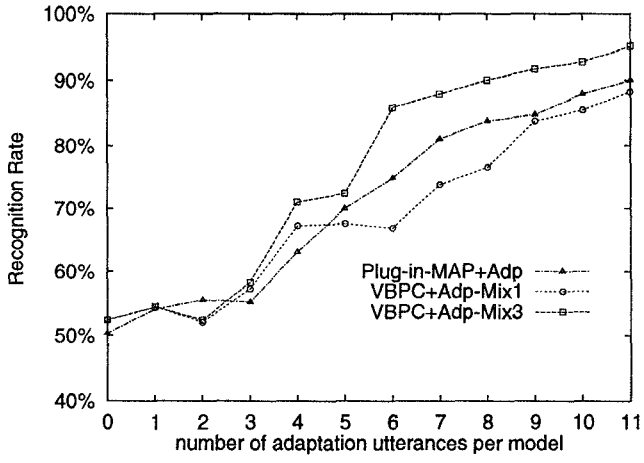

Figure 2: Performance comparison of cross-gender speech recognition as a function of amount of adaptation data among methods in which sequential Bayesian learning is combined with plug-in MAP decoding or VBPC (with mixture number $M = 1, 3$)

the condition of unlimited memory and calculation, some suboptimal methods are needed in practice. The N-Best implementation studied here is sensitive to mixands pruning, selection, and merging in the sequential adaptation procedure.

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 Speech and Audio Processing*, Vol. 5, No. 2, 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* (Munich, Germany), April 1997, 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* (Rhodes, Greece), September 1997, pp.1847-1850.

[4] Q. Huo and C.-H. Lee, "A Bayesian predictive classification approach to robust speech recognition," submitted to *IEEE Trans. on Speech and Audio Processing*, August 1997.

[5] H. Jiang, K. Hirose and Q. Huo, "Robust speech recognition based on Bayesian prediction approach", submitted to *IEEE Trans. on Speech and Audio Processing*, August 1997. See also a condensed version titled "Robust speech recognition based on Viterbi Bayesian predictive classification," in *Proc. ICASSP-97*, pp.II-1551-1554, 1997.