Effects of Bayesian predictive classification using variational Bayesian posteriors for sparse training data in speech recognition

Shinji Watanabe and Atsushi Nakamura

NTT Communication Science Laboratories, NTT Corporation

{watanabe, ats}@cslab.kecl.ntt.co.jp

Abstract

We introduce a robust classification method using Bayesian predictive distribution (Bayesian predictive classification, referred to as BPC) into speech recognition. We and others have recently proposed a total Bayesian framework for speech recognition, Variational Bayesian Estimation and Clustering for speech recognition (VBEC). VBEC includes an analytical derivation of approximate posterior distributions that are essential for BPC, based on variational Bayes (VB). BPC using VB posterior distributions (VB-BPC) can mitigate the over-training effects by marginalizing output distribution. We address the sparse data problem in speech recognition, and show how VB-BPC is robust against the data sparseness, experimentally.

1. Introduction

The performance of statistical speech recognition is greatly degraded when it encounters unseen environments. This is because speech recognition is based on the conventional Maximum Likelihood (ML) approaches, which often over-train model parameters to fit limited amount of training data (over-training problem). On the other hand, a Bayesian framework can mitigate the effects of the over-training problem because of the following three important advantages: effective utilization of prior knowledge, appropriate selection of model structure and robust classification of unseen speech, each of which works to mitigate the effects of over training. Recently, we and others proposed Variational Bayesian Estimation and Clustering for speech recognition (VBEC), which includes all the above Bayesian advantages [1]. In this paper, we focus on the third Bayesian advantage, the robust classification of unseen speech.

Using a conventional classification method based on the ML approach (MLC), we prepare a probability function (for example $f(x;\theta)$ with a model parameter θ), which represents the distribution of features for a classification category, e.g. a phoneme category, and point-estimate the parameter θ using labeled speech data. However, the parameter is often estimated incorrectly because of the sparseness of the training data and the mismatch between the training and input data. Therefore, MLC might be seriously affected by incorrectness in estimation when classifying unseen speech data. In contrast, a classification method based on the Bayesian approach does not use the point-estimated value of the parameter, but assumes that the value itself also has a probability distribution represented by a function (for example $g(\theta)$). Then, by taking the expectation of $f(x;\theta)$ with respect to $g(\theta)$, we obtain a distribution of x, which can robustly predict the behavior of unseen data. Some previous studies proposed classification methods based on the predictive distribution (Bayesian Predictive Classification, referred to as BPC) for speech recognition, and proved that they were capable of much more robust classification than MLC [2, 3].

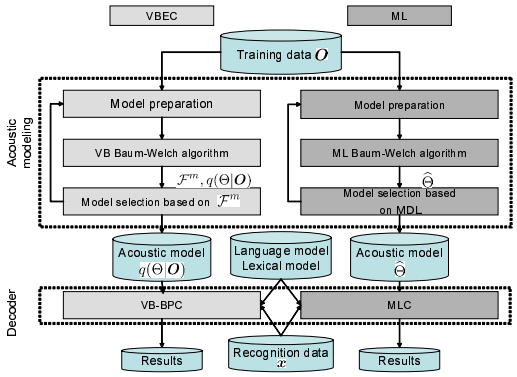

Figure 1: VBEC framework.

A major problem with BPC is how to provide $g(\theta)$. In [2, 3], they prepare $g(\theta)$ as a prior distribution, and incorporate the knowledge from input speech into BPC through the hyper-parameters of $g(\theta)$ to accommodate the mismatch between training data and input speech. Another point of interest involves finding a way to realize BPC using only training data. In fact, it is very natural for VBEC to provide $g(\theta)$ from training data since the VBEC framework is designed to deal consistently with a posterior distribution of θ, which is a direct realization of $g(\theta)$, by variational Bayes (VB-posteriors) [4]. Consequently we examine the effect of BPC using VB-posteriors (VB-BPC) on the sparse data problem, and show the effectiveness of VB-BPC through speech recognition experiments.

2. VB posteriors within VBEC framework

First, we briefly describe the VBEC framework by providing definitions of the VB posteriors, which are used in Section 3, where we explain VB-BPC. VBEC is a total Bayesian framework in the sense that all its training and classification procedures are consistently based on approximated posteriors by using VB (as shown in Figure 1 and see [1] for details).

VBEC is formulated in the standard acoustic models of speech recognition, which are represented by the Context-Dependent Hidden Markov Models (CDHMM) and the multivariate Gaussian Mixture Models (GMM). In the standard acoustic models, the output distributions of t-th frame training data $\boldsymbol{O}^t \in \mathcal{R}^D$ at the HMM state transition from i to j in a phoneme category c is represented as follows:

$$p(\boldsymbol{O}^t|c,i,j,\Theta) = a_{ij}\sum_k w_{jk}\mathcal{N}(\boldsymbol{O}^t|\boldsymbol{\mu}_{jk},\Sigma_{jk}), \qquad (1)$$

a_{ij} denotes the state transition probability from state i to state j, and w_{jk} denotes the k-th weight factor of the Gaussian mixture for state j. $\mathcal{N}(\boldsymbol{O}^t|\boldsymbol{\mu}_{jk},\Sigma_{jk})$ denotes a Gaussian with

a mean vector $\boldsymbol{\mu}_{jk}$ and covariance matrix Σ_{jk}. Therefore, output distributions of acoustic models are parameterized by $\Theta_{ij} = \{a_{ij}, w_{jk}, \boldsymbol{\mu}_{jk}, \Sigma_{jk}^{-1} | k = 1, ...\}$. In this section, the phoneme category index c is excluded to avoid a complicated equation.

We can obtain the VB posterior distributions for the model parameters $q(\Theta|\boldsymbol{O})$, from a VB calculation by using the output distribution (Eq. (1)) and the conjugate prior distributions. We summarize the calculated results of the VB posterior distributions for the model parameters as follows:

$$q(\Theta|\boldsymbol{O}) = \prod_i \mathcal{D}(\{a_{ij}\}_j|\{\phi_{ij}\}_j) \prod_j \mathcal{D}(\{w_{jk}\}_k|\{\varphi_{jk}\}_k)$$
$$\prod_k \mathcal{N}(\boldsymbol{\mu}_{jk}|\boldsymbol{\nu}_{jk}, (\xi_{jk})^{-1}\Sigma_{jk}) \prod_d \mathcal{G}(\Sigma_{jk,d}^{-1}|\eta_{jk}, R_{jk,d}). \quad (2)$$

In Eq. (2), $\mathcal{D}$ denotes a Dirichlet distribution and $\mathcal{G}$ denotes a gamma distribution. The posterior distributions of a_{ij} and w_{jk} are represented by Dirichlet distributions, and the posterior distribution of $\boldsymbol{\mu}_{jk}$ and Σ_{jk} is represented by a normal-gamma distribution. If the covariance matrix elements are off the diagonal, a normal-Wishart distribution is set as the posterior distribution of $\boldsymbol{\mu}_{jk}$ and Σ_{jk}. These posterior distributions were parameterized by hyper-parameters defined as:

$$\begin{cases} \phi_{ij} = \phi_{ij}^0 + \gamma_{ij}, \qquad \varphi_{jk} = \varphi_{jk}^0 + \zeta_{jk}, \quad \xi_{jk} = \xi_{jk}^0 + \zeta_{jk} \\ \boldsymbol{\nu}_{jk} = \dfrac{\xi_{jk}^0 \boldsymbol{\nu}_{jk}^0 + \boldsymbol{\kappa}_{jk}}{\xi_{jk}^0 + \zeta_{jk}}, \eta_{jk} = \eta_{jk}^0 + \zeta_{jk} \\ R_{jk} = R_{jk}^0 + \Xi_{jk} - \dfrac{1}{\zeta_{jk}} \boldsymbol{\kappa}_{jk}(\boldsymbol{\kappa}_{jk})' \\ \qquad + \dfrac{\xi_{jk}^0 \zeta_{jk}}{\xi_{jk}^0 + \zeta_{jk}} \left(\dfrac{\boldsymbol{\kappa}_{jk}}{\zeta_{jk}} - \boldsymbol{\nu}_{jk}^0 \right) \left(\dfrac{\boldsymbol{\kappa}_{jk}}{\zeta_{jk}} - \boldsymbol{\nu}_{jk}^0 \right)' \end{cases} \quad (3)$$

where γ_{ij}, ζ_{jk}, $\boldsymbol{\kappa}_{jk}$ and Ξ_{jk} denotes 0th, 1st and 2nd order sufficient statistics, respectively. $\phi^0, \varphi^0, \xi^0, \boldsymbol{\nu}^0, \eta^0$ and R^0 are hyper-parameters of prior distributions and are set in the initial training phase.

The sufficient statistics $\gamma_{ij}, \zeta_{jk}, \boldsymbol{\kappa}_{jk}$ and Ξ_{jk} are calculated by using γ_{ij}^t and ζ_{jk}^t as follows:

$$\gamma_{ij} = \sum_t \gamma_{ij}^t, \qquad \zeta_{jk} = \sum_t \zeta_{jk}^t$$
$$\boldsymbol{\kappa}_{jk} = \sum_t \zeta_{jk}^t \boldsymbol{O}^t, \quad \Xi_{jk} = \sum_t \zeta_{jk}^t \boldsymbol{O}^t (\boldsymbol{O}^t)'. \quad (4)$$

γ_{ij}^t is a VB transition posterior distribution, which denotes the transition probability from a state i to a state j at a frame t, and ζ_{jk}^t is a VB occupation posterior distribution, which denotes the occupation probability of a mixture component k in a state j at a frame t, in the VB approach. These are computed efficiently by using a probabilistic assignment via the familiar forward-backward algorithm or a deterministic assignment via the Viterbi algorithm.

Thus, the VB posteriors can be calculated by computing the VB forward-backward or Viterbi algorithm, and updating the VB posteriors iteratively. This algorithm operates in the same way as the Baum-Welch algorithm based on the ML approach, and we refer to these calculations as a VB Baum-Welch algorithm, which is proposed in [1]. VBEC is based on the VB Baum-Welch algorithm. As with the likelihood in the ML framework, VBEC also has an objective function to judge the convergence of the VB Baum-Welch algorithm using the VBEC objective function $\mathcal{F}^m$. The VBEC objective function plays a double role, i.e., VBEC is not only a criterion with which to judge the convergence of the VB Baum-Welch algorithm, but also a criterion for determining the model topology. The VBEC objective function, which was first derived in [1], accurately reflects all the topological effects of the standard acoustic models based on the VB framework. Therefore, we can select the optimal model topology that maximizes the VBEC objective function. Thus, the acoustic modeling of the VBEC framework outputs the appropriate VB posteriors $q(\Theta|\boldsymbol{O})$ based on the appropriate model topology, as shown in Figure 1.

3. VB-BPC

3.1. Formulation

In this section, we formulate BPC by explicitly distinguishing between training data $\boldsymbol{O}$ and input data $\boldsymbol{x}$. After the acoustic modeling described in Section 2, we obtain the appropriate VB posterior distributions $q(\Theta|\boldsymbol{O})$ for the appropriate model structure. In recognition, an input speech $\boldsymbol{x}^t$ for a frame t is classified as the optimal phoneme class $\bar{c}$ using $p(c|\boldsymbol{x}^t, \boldsymbol{O})$ defined as follows:

$$\bar{c} = \arg\max_c p(c|\boldsymbol{x}^t, \boldsymbol{O}) \cong \arg\max_c p(c) p(\boldsymbol{x}^t|c, \boldsymbol{O}). \quad (5)$$

Here, $p(c)$ is the class prior distribution obtained by language and lexicon models, and $p(\boldsymbol{x}^t|c, \boldsymbol{O})$ is the predictive posterior distribution. c is assumed to be independent of $\boldsymbol{O}$ (i.e., $p(c|\boldsymbol{O}) \cong p(c)$). We focus on the predictive posterior distribution $p(\boldsymbol{x}^t|c, \boldsymbol{O})$, and consider $\boldsymbol{x}^t$ to be in a particular state j of a category c, which has transited from a previous state i. Then, by introducing the distribution parameter Θ, $p(\boldsymbol{x}^t|c, i, j, \boldsymbol{O})$ is obtained as follows:

$$\begin{aligned} p(\boldsymbol{x}^t|c, i, j, \boldsymbol{O}) &= \int p(\boldsymbol{x}^t|c, i, j, \Theta) p(\Theta|c, i, j, \boldsymbol{O}) d\Theta \\ &= \int p(\boldsymbol{x}^t|\Theta_{ij}^{(c)}) p(\Theta_{ij}^{(c)}|\boldsymbol{O}) d\Theta_{ij}^{(c)} \end{aligned} \quad (6)$$

where $\Theta_{ij}^{(c)} \equiv \{a_{ij}^{(c)}, w_{jk}^{(c)}, \boldsymbol{\mu}_{jk}^{(c)}, \Sigma_{jk}^{(c)} | k = 1, ...\}$ is a set of model parameters in a category c. Therefore, by calculating the integral in Eq. (6), a sequence score can be computed by summing up each frame score based on the Viterbi algorithm that enables input speech to be classified. The approach that involves considering the integrals and true posterior distributions in Eq. (6) is called the Bayesian inference or Bayesian Predictive Classification (BPC) approach. However, in general, a true posterior distribution $p(\Theta_{ij}^{(c)}|, \boldsymbol{O})$ is difficult to obtain analytically since the integral in Eq. (6) is complicated. On the other hand, the numerical approach requires a very long computation time and is unrealistic for use in speech recognition.

Conventional ML-based Classification (MLC) is considered to be an approximation of BPC. With the ML approach, the predictive posterior distribution is calculated by approximating $p(\Theta_{ij}^{(c)}|\boldsymbol{O})$ as a *Dirac delta function* of ML estimates $\widehat{\Theta}_{ij}^{(c)} \equiv \{\widehat{a}_{ij}^{(c)}, \widehat{w}_{jk}^{(c)}, \widehat{\boldsymbol{\mu}}_{jk}^{(c)}, \widehat{\Sigma}_{jk}^{(c)} | k = 1, ...\}$, instead of using the true posterior distribution, as follows [1]:

$$\begin{aligned} p(\boldsymbol{x}^t|c, i, j, \boldsymbol{O}) &\cong \int p(\boldsymbol{x}^t|\Theta_{ij}^{(c)}) \delta(\Theta_{ij}^{(c)} - \widehat{\Theta}_{ij}^{(c)}) d\Theta_{ij}^{(c)} \\ &= p(\boldsymbol{x}^t|\widehat{\Theta}_{ij}^{(c)}) = \widehat{a}_{ij}^{(c)} \sum_k \widehat{w}_{jk}^{(c)} \mathcal{N}(\boldsymbol{x}^t|\widehat{\boldsymbol{\mu}}_{jk}^{(c)}, \widehat{\Sigma}_{jk}^{(c)}) \end{aligned} \quad (7)$$

[1] In this paper, we also deal with the classification using MAP estimates instead of using the ML estimates as MLC.

where $\delta(y-y')$ is a delta function defined as $\int g(y)\delta(y-y')dy = g(y')$. Therefore, from Eq. (7), MLC is based on the Gaussians. There are two main problems with MLC, both of which degrade the classification performance. First, if the training data is sparse, the approximation in Eq. (7) is ineffective, and the classification performance degrades (sparse data problem). Second, MLC depends on ML estimates $\widehat{\Theta}$, which are obtained simply from training data, and therefore, MLC cannot deal with mismatches between training and testing conditions. This also leads to a degradation in the classification performance (mismatch problem).

In contrast, by using the estimated VB posterior distributions $q(\Theta_{ij}^{(c)}|\boldsymbol{O})$, VBEC could consider the integrals in Eq. (6) and mitigate the MLC problems. When we approximate the true posterior distribution $p(\Theta_{ij}^{(c)}|\boldsymbol{O})$ by using the estimated VB posterior distributions $q(\Theta_{ij}^{(c)}|\boldsymbol{O})$, the integral over $\Theta_{ij}^{(c)}$ can be solved analytically by Eqs. (1) and (2) and is found to be a Student mixture distribution, as follows:

$$
\begin{aligned}
p(\boldsymbol{x}^t|c,i,j,\boldsymbol{O}) &\cong \int d\Theta_{ij}^{(c)} p(\boldsymbol{x}^t|\Theta_{ij}^{(c)}) q(\Theta_{ij}^{(c)}|\boldsymbol{O}) \\
&= \frac{\phi_{ij}^{(c)}}{\sum_j \phi_{ij}^{(c)}} \sum_k \frac{\varphi_{jk}^{(c)}}{\sum_k \varphi_{jk}^{(c)}} \\
&\quad \prod_d \mathcal{S}t\left(\boldsymbol{x}_d^t \middle| \boldsymbol{\nu}_{jk,d}^{(c)}, \frac{1+\xi_{jk}^{(c)}}{\xi_{jk}^{(c)}} R_{jk,d}^{(c)}, \eta_{jk}^{(c)}\right) \qquad (8)
\end{aligned}
$$

Here, $\mathcal{S}t(\cdot)$ is the Student distribution. Therefore, input speech can be classified by using the predictive score obtained from Eq. (8). We call this approach Bayesian Predictive Classification using VB posterior distributions (VB-BPC). VB-BPC achieves VBEC with a total Bayesian framework for speech recognition that possesses a consistent concept whereby all procedures (acoustic modeling and speech classification) are carried out based on posterior distributions, as shown in Figure 1. VBEC mitigates the "sparse data problem" by using the full potential of the Bayesian approach that is drawn out by the consistent concept, and there, the VB-BPC contributes greatly as one of the components.

There are other BPC techniques that can effectively mitigate the model "mismatch problem" [2, 3]. They incorporate the knowledge from input speech into BPC through prior distribution, and accommodate the mismatch between training and input data. On the other hand, VB-BPC, in this paper, is purely derived from training data and does not use input data, since we intend to evaluate the acoustic model robustness derived from only training data. Therefore, VB-BPC is completely open for input data, which might lose the mitigation effect of the mismatch problem. For the sparse data problem, VB-BPC does not require an asymptotic condition based on sufficient training data, unlike [2, 3]. Therefore, the function form of the analytical result is a Student distribution, which is different from the Gaussian based results obtained by [2, 3] based on the Laplace approximation.

3.2. Student distribution

One of the specifications of VB-BPC is that its classification function is represented by a non-Gaussian Student distribution. Here, we discuss the Student distribution to clarify the difference between Gaussian and Student distributions. The Student distribution is defined as follows:

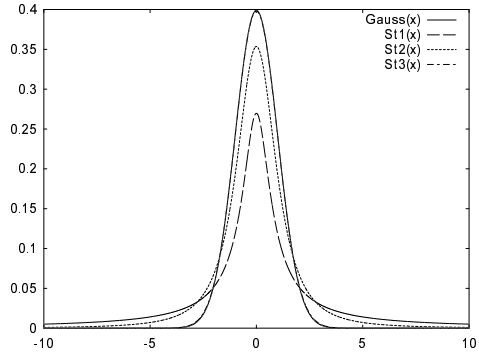

Figure 2: Gaussian (Gauss(x)) and three Student distributions (St1(x), St2(x) and St3(x)). The mean and variance parameter of the Gaussian and Student distributions are the same. Only the degrees of freedom (DoF) of the Student distributions are changed (DoF values of St1, St2 and St3 are 0.5, 2 and 100)

$$
\begin{aligned}
&\mathcal{S}t(x|\rho,\lambda,\alpha) \\
&= \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)\Gamma\left(\frac{1}{2}\right)} \left(\frac{\lambda}{\alpha}\right)^{\frac{1}{2}} \left(1+\frac{\lambda}{\alpha}(x-\rho)^2\right)^{-\frac{\alpha+1}{2}}. \qquad (9)
\end{aligned}
$$

Here ρ and λ correspond to the mean and variance of the Gaussian, respectively. The Student distribution has an additional parameter α, which is referred to as a degree of freedom. This parameter represents the wideness of the distribution as shown in Figure 2. If α is small, the distribution becomes wider than the Gaussian, and if α is large, it approaches the Gaussian. From Eq. (3), $\alpha = \eta_{jk}$ is proportional to the training data occupation counts. Therefore, with dense training data, $\alpha = \eta_{jk}$ becomes large and VB-BPC becomes similar to the Gaussian-based MLC. On the other hand, when the training data is sparse, $\alpha = \eta_{jk}$ becomes small, and the distribution becomes wider. This behavior is effective for the sparse training data problem because the wider distribution could cover the difference between the input data and the biased training data due to the sparseness. Consequently VB-BPC mitigates the effects of the sparse training data problem.

4. Experiments

We used experiments to examine the effectiveness of VB-BPC as a solution to the sparse training data problem. We conducted isolated word recognition experiments and compared the fully implemented version of VBEC that includes VB-BPC

Table 1: Experimental conditions for isolated word recognition

Sampling rate/quantization	16 kHz / 16 bit
Feature vector	12 order MFCC with energy
(39 dimensions)	+Δ+ΔΔ (CMN)
Window	Hamming
Frame size/shift	25/10 ms
Num. of states	3 (Left to right)
Num. of phoneme categories	27
Num. of phonetic questions	44
Training data	ASJ: 3,000 utterances, 4.1 hours (male)
Test data	JEIDA: 100 city names, 7,200 words (male)

ASJ: Acoustical Society of Japan

JEIDA: Japan Electronic Industry Development Association

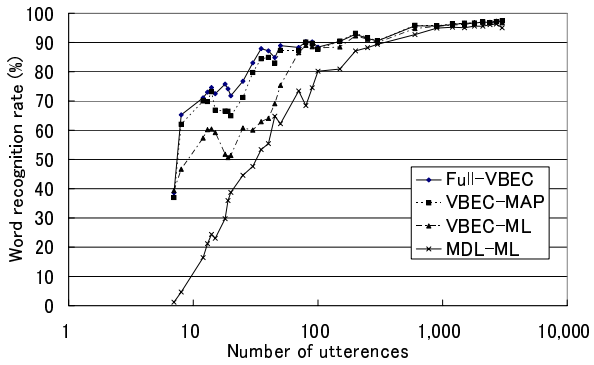

Figure 3: Recognition rate for varying the amount of training data.

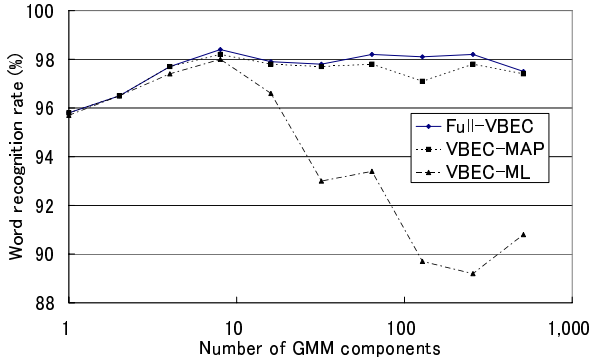

Figure 4: Recognition rate for varying the number of GMM components.

with other partially implemented versions. The experimental conditions are summarized in Table 1. The training data consisted of about 3,000 Japanese utterances (4.1 hours) spoken by 30 males. The test data consisted of 100 Japanese city names spoken by 75 males (a total of 7,200 words). Several subsets of difference sizes were randomly extracted from the training data set, and each of the subsets was used to construct a set of acoustic models. The acoustic models were represented by CDHMMs and each HMM state has a 16 component Gaussian mixture. As a result, 34 sets of acoustic models for various amounts of training data were prepared.

Table 2 summarizes approaches that use a combination of methods for model selection (or HMM state clustering), training and classification, for each of which we employ either VB or other approaches. The combination determines how well it includes the Bayesian advantages, i.e., effective utilization of prior knowledge, appropriate selection of model structure and robust classification of unseen speech. Here MDL indicates a model selection using the minimum description length criterion, which we should recognize as a kind of ML-based approach [5], and MAP means a classification using a function estimated based on the maximum a posteriori criterion, which is regarded as a partial implementation of the Bayesian approach [6]. Note that all of the combinations, excepting for Full-VBEC, include an ML or a merely partial implementation of the Bayesian approach, and that the approaches are listed in the order of how well the Bayesian advantages are included. Figures 3 shows recognition results obtained using the combinations. We can see that the better the Bayesian advantages were included, the more robustly speech was recognized. In particular, in cases of sparse training data (less than 100 utterances), Full-VBEC significantly outperformed the other combinations. In addition, when the training data is more sparse (less than 50 utterances), Full-VBEC was better than VBEC-MAP by $2 \sim 9$ %. Note that the only difference between them was the classification algorithm, i.e., VB-BPC or MAP. This improvement is obviously due to the effectiveness of VB-BPC, and perhaps due to the synergistic effect that results from exploiting the full potential of the Bayesian approach by incorporating all its advantages.

To support these results, we also examined the VB-BPC effects for another aspect of the sparse training data. In experiments using training data that consisted of 3,000 utterances, we varied the numbers of Gaussian components per state from 1 to 512. The structure of the CDHMMs determined by VBEC was fixed and we compared Full-VBEC, VBEC-MAP and VBEC-ML for the same CDHMM structure. In this case, the number of training data per parameter decreases as the number of components increases, and the sparse data problem also occurs when there is a large number of components. Figure 4 shows the word recognition rate for various numbers of Gaussians per state. Full-VBEC outperformed VBEC-MAP and VBEC-ML, especially when there were large numbers of Gaussians per state. Thus, this and the previous experiments confirm the effectiveness of the VB-BPC for the sparse data problem.

5. Summary

In this paper, we introduce a method of Bayesian Predictive Classification using Variational Bayesian posteriors (VB-BPC) in speech recognition. The effect for the sparse data problem is confirmed by the recognition rate improvement using VB-BPC, experimentally. VB-BPC also has the potential to mitigate the mismatch problems between training and input data that originate in such factors as speaker variations, environment noise and speaking styles. We plan to examine the VB-BPC effect in relation to such mismatch problems in the future.

6. References

[1] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Variational Bayesian estimation and clustering for speech recognition," *IEEE Trans. SAP*, vol. 12, pp. 365–381, 2004.

[2] Q. Huo and C-H. Lee, "A Bayesian predictive classification approach to robust speech recognition," *IEEE Trans. SAP*, vol. 8, pp. 200–204, 2000.

[3] H. Jiang, K. Hirose, and Q. Huo, "Robust speech recognition based on a Bayesian prediction approach," *IEEE Trans. SAP*, vol. 7, pp. 426–440, 1999.

[4] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," in *Proc. UAI 15*, 1999.

[5] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in *Proceedings of EuroSpeech1997*, 1997, vol. 1, pp. 99–102.

[6] J-L. Gauvain and C-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. SAP*, vol. 2, pp. 291–298, 1994.

Table 2: Configuration of VBEC and ML based approaches

	Model selection	Training	Classification
Full-VBEC	**VB**	**VB**	**VB-BPC**
VBEC-MAP	**VB**	**VB**	MAP
VBEC-ML	**VB**	ML	MLC
MDL-ML	MDL	ML	MLC