Bayesian Approaches in Speech Recognition

Shinji Watanabe*
* NTT Communication Science Laboratories, NTT Coporation, Kyoto, Japan
E-mail: watanabe.shinji@lab.ntt.co.jp

***Abstract*—This paper focuses on applications of Bayesian approaches to speech recognition. Bayesian approaches have been widely studied in statistics and machine learning fields, and one of the advantages of the Bayesian approaches is to improve generalization ability compared to maximum likelihood approaches. The effectiveness for speech recognition is shown experimentally in speaker adaptation tasks by using Maximum A Posterior (MAP) and model complexity control by using Bayesian Information Criterion (BIC). This paper introduces the variational Bayesian approaches, in addition to the MAP, BIC and other Bayesian techniques, for speech recognition. VBEC (Variational Bayesian Estimation and Clustering for speech recognition) is a fully Bayesian speech recognition framework, and achieves robust acoustic modeling and speech classification. This paper explains the formulation and experimental effectiveness of these Bayesian approaches for speech recognition.**

I. Introduction

Speech recognition, which converts speech information into text information, is the core technology for allowing computers to understand the human intent. The current successes in speech recognition are based on pattern recognition, which uses statistical learning theory. Maximum Likelihood (ML) methods have become the standard techniques for constructing acoustic and language models for speech recognition. ML methods guarantee that ML estimates approach the true values of the parameters. ML methods have been used in various aspects of statistical learning, and especially for acoustic modeling in speech recognition since the Expectation-Maximization (EM) algorithm [1] is a practical way of obtaining the local optimum solution for the training of latent variable models. Therefore, acoustic modeling based on Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) have been developed greatly by using the ML-EM approach [2]–[4].

However, the performance of current speech recognition systems is far from satisfactory. Specifically, the recognition performance is much poorer than the human recognition ability since speech recognition suffers from a distinct lack of robustness to unknown conditions, which is crucial for practical use. In a real environment, there are many fluctuations originating from various factors such as the speaker, context, speaking style and noise. In fact, the performance of acoustic models trained using read speech decreases greatly when the models are used to recognize spontaneous speech due to the mismatch between the read and spontaneous speech environments [5]. Therefore, most of the problems posed by current speech recognition techniques result from a lack of robustness. This lack of robustness is an obstacle to the deployment of commercial applications based on speech recognition technology, and improving robustness has been a common worldwide challenge in the field of acoustic and language studies. Acoustic studies have taken mainly two directions: the improvement of acoustic models beyond the conventional HMM, and the improvement of the acoustic model learning method beyond the conventional ML approach. This paper addresses the challenge in terms of improving the learning method by employing *Bayesian* approaches.

In Bayesian approaches, all the variables introduced when models are parameterized, such as model parameters and latent variables, are regarded as probabilistic variables, and their posterior distributions are obtained based on the Bayes rule. The difference between the Bayesian and ML approaches is that the target of estimation is a *distribution function* in the Bayesian approach whereas it is a *parameter value* in the ML approach. Based on this posterior distribution estimation, the Bayesian approach can generally achieve more robust model construction and classification than an ML approach [6], [7]. In fact, the Bayesian approach has the following three advantages:

(A) Effective utilization of prior knowledge through prior distributions (prior utilization)
(B) Model selection that obtains a model structure with the highest probability of posterior distribution of model structures (model selection)
(C) Robust classification by marginalizing model parameters (robust classification)

In general, these advantages make a pattern recognition method more robust than that based on the ML approaches.

However, the Bayesian approach requires complex integral and expectation computations to obtain posterior distributions when models have latent variables. The current acoustic model in speech recognition has the latent variables included in an HMM and a Gaussian Mixture Model (GMM). Therefore, some approximated Bayesian techniques are applied to speech recognition to avoid the computational problem. For example, the Maximum A Posteriori based framework approximates the posterior distribution of the parameter by using the MAP approximation to utilize prior information [8], [9]. Bayesian Information Criterion (BIC) [10]–[13][1] and Bayesian Predictive Classification (BPC) [14], [15] based frameworks partially realize Bayesian advantages for model selection and robust classification, respectively, in speech recognition. These ap-

[1] BIC and Minimum Description Length (MDL) criterion have been independently proposed, but they are practically the same. Therefore, they are identified in this paper and referred to as BIC/MDL.

proaches are simple and powerful frameworks to realize the Bayesian advantages in speech recognition. However, these approaches lose a part of the Bayesian advantages due to their approximation, as shown in Table I.

Therefore, this paper introduces a speech recognition framework based on a fully Bayesian approach to overcome the lack of robustness described above by utilizing the three Bayesian advantages [16], [17]. Recently, a *Variational Bayesian* (VB) approach was proposed in the learning theory field that avoids complex computations by employing the variational approximation technique [18]–[21]. With this VB approach, approximate posterior distributions (VB posterior distributions) can be obtained effectively by iterative calculations similar to the Expectation-Maximization algorithm used in the ML approach, while the three advantages of the Bayesian approaches are still retained. Therefore, the framework is formulated using VB to replace the ML approaches with the Bayesian approaches in speech recognition. The proposed Variational Bayesian Estimation and Clustering for speech recognition (VBEC) is a *fully Bayesian framework,* where all acoustic procedures for speech recognition (acoustic model construction and speech classification) are based on the VB posterior distribution. Consequently, VBEC includes the three Bayesian advantages unlike the conventional Bayesian approaches, as shown in Table I. This paper also confirms experimentally the effectiveness of the three Bayesian advantages, prior utilization, model selection and robust classification, in VBEC.

II. Approximated Bayesian approaches

In this section, we briefly review the Bayesian approach in contrast with the ML approach, and introduce approximated Bayesian approaches, which are widely used for speech recognition.

A. Maximum A Posteriori (MAP) approaches

MAP approaches are introduced to speech recognition to utilize the prior information of the Bayesian advantages [8], [9]. The Bayesian approach is based on posterior distributions, while the ML approach is based on distribution parameters. Let $\mathbf{O} \in \{\mathbf{o}_t \in \mathbb{R}^D | t = 1, \cdots, T\}$ be a given training data set of D dimensional feature vectors and $\mathbf{Z} \in \{z_t | t = 1, \cdots, T\}$ be a set of the corresponding latent variables. The posterior distribution for a distribution parameter Θ_c of category c is obtained by the famous Bayes theorem as follows:

$$p(\Theta_c|\mathbf{O}, m) = \sum_{\mathbf{Z}} \int \frac{p(\mathbf{O}, \mathbf{Z}|\Theta, m)p(\Theta|m)}{p(\mathbf{O}|m)} d\Theta_{-c}, \qquad (1)$$

TABLE I
Comparison of VBEC and other Bayesian frameworks in terms of Bayesian advantages

Bayesian advantage	VBEC	MAP	BIC/MDL	BPC
(A) Prior utilization	√	√	–	–
(B) Model selection	√	–	√	–
(C) Robust classification	√	–	–	√

where $p(\Theta|m)$ is a prior distribution for all distribution parameters Θ, and m denotes the model structure index, for example, the number of Gaussian components or HMM states. Here, $-c$ represents the set of all categories without c. In this paper, we regard the hyper-parameter setting as the model structure, and include its variations in index m. From Eq. (1), prior information can be utilized via estimations of the posterior distribution, which depends on prior distributions.

The calculation of Eq. (1) cannot be solved analytically due to the summation over latent variables. To avoid the problem, the MAP approaches approximate the distribution estimation to the point estimation. Namely, instead of obtaining the posterior distribution in Eq. (1), the MAP approach obtains the following value

$$\begin{aligned} \Theta_c^{MAP} &= \underset{\Theta_c}{\text{argmax}}\, p(\Theta_c|\mathbf{O}, m) \\ &= \sum_{\mathbf{Z}} p(\mathbf{O}, \mathbf{Z}|\Theta_c, m)p(\Theta_c|m). \end{aligned} \qquad (2)$$

This estimation can be efficiently performed by using the EM algorithm. The MAP approximation is first applied to the estimation of single-Gaussian HMM parameters in [8] and is extended to GMM-HMMs in [9]. The effectiveness of the MAP approach can be shown in a speaker recognition task where prior distributions are set by speaker-independent HMMs.

B. Bayesian Information Criterion (BIC) approaches

BIC approaches are introduced to speech recognition to perform model selection [10], [11], To deal with model structure in a Bayesian approach, we can consider the following posterior distribution:

$$p(m|\mathbf{O}) = \sum_{\mathbf{Z}} \int \frac{p(\mathbf{O}, \mathbf{Z}|\Theta, m)p(\Theta|m)p(m)}{p(\mathbf{O})} d\Theta, \qquad (3)$$

where $p(m)$ denotes a prior distribution for model structure m. However, similar to the MAP approach, the calculation of Eq. (3) cannot be solved analytically due to the summation over latent variables. The Bayesian Information Criterion (BIC) only focuses on the model that does not have latent variables, and can obtain the following equation based on the asymptotic assumption (large amount of data assumption):

$$\log p(m|\mathbf{O}) \propto \log p(\mathbf{O}|\Theta, m) - \frac{\#(\Theta)}{2} \log T. \qquad (4)$$

The first term in the right hand side is a log-likelihood term and the second term is a penalty term, which is proportional to the number of model parameters ($\#(\Theta)$).

This criterion is widely used for speech processing (e.g., phonetic decision tree clustering [10], [11], speaker clustering [12], and Gaussian pruning in acoustic models [13]).

C. Bayesian Predictive Classification (BPC) approaches

To perform robust classification, BPC approaches are used in speech recognition [14], [15]. Once the posterior distribution $p(\Theta_c|\mathbf{O}, m)$ is estimated for all categories, the category for test data $\mathbf{x}$ is determined by:

$$\bar{c} = \underset{\{c\}}{\text{argmax}} \int p(\mathbf{x}|\Theta, c, m) p(\Theta|c, \mathbf{O}, m) d\Theta. \tag{5}$$

The parameters are integrated out in Eq. (5) so that the effect of over-training is mitigated, and robust classification is obtained. The approach that involves considering the integrals and true posterior distributions in Eq. (5) is called the Bayesian Predictive Classification (BPC) approach. In general, a true posterior distribution $p(\Theta|c\mathbf{O}, m)$ is difficult to obtain analytically. [14] and [15] uses posterior distributions, which have mean parameters on point-estimated model parameters (e.g. ML or MAP estimates) and variance parameters as control parameters.

Thus, MAP, BIC, and BPC approaches can be practically realized in speech recognition, having three Bayesian advantages, respectively. Next section, we introduce Variational Bayesian Estimation and Clustering for speech recognition (VBEC), which includes the three Bayesian advantages at the same time unlike the MAP, BIC, and BPC approaches, as shown in Table I.

III. Variational Bayesian Estimation and Clustering for speech recognition

A. Variational Bayesian approach

We briefly explain a variational Bayesian approach [18]–[21]. To begin with, we assume that

$$q(\Theta, \mathbf{Z}|\mathbf{O}, m) = \prod_c q(\Theta_c|\mathbf{O}_c, m) q(\mathbf{Z}_c|\mathbf{O}_c, m), \tag{6}$$

where q is an arbitrary posterior distribution. Then, the variational Bayes focuses on the following objective functional

$$\begin{aligned} &\mathcal{F}^m[q(\Theta|\mathbf{O}, m), q(\mathbf{Z}|\mathbf{O}, m)] \\ &= \left\langle \log \frac{p(\mathbf{O}, \mathbf{Z}|\Theta, m) p(\Theta|m)}{q(\Theta|\mathbf{O}, m) q(\mathbf{Z}|\mathbf{O}, m)} \right\rangle_{q(\Theta|\boldsymbol{O}, m), q(\mathbf{Z}|\boldsymbol{O}, m)}. \end{aligned} \tag{7}$$

Here, the brackets $\langle\rangle$ denote the expectation i.e. $\langle g(y) \rangle_{p(y)} \equiv \int g(y) p(y) dy$ for a continuous variable y and $\langle g(n) \rangle_{p(n)} \equiv \sum_n g(n) p(n)$ for a discrete variable n. Eq. (7) is a lower bound of marginalized log likelihood. Therefore, the optimal posterior distribution can be obtained by a variational method, which results in maximizing the functional $\mathcal{F}$, i.e.,

$$\begin{aligned} \widetilde{q}(\Theta_c|\mathbf{O}, m) &= \underset{q(\Theta_c|\mathbf{O}, m)}{\text{argmax}} \ \mathcal{F}^m[q(\Theta|\mathbf{O}, m), q(\mathbf{Z}|\mathbf{O}, m)]. \\ \widetilde{q}(\mathbf{Z}_c|\mathbf{O}, m) &= \underset{q(\mathbf{Z}_c|\mathbf{O}, m)}{\text{argmax}} \ \mathcal{F}^m[q(\Theta|\mathbf{O}, m), q(\mathbf{Z}|\mathbf{O}, m)]. \\ \widetilde{q}(m|\mathbf{O}) &= \underset{q(m|\mathbf{O})}{\text{argmax}} \mathcal{F}^m[q(\Theta|\mathbf{O}, m), q(\mathbf{Z}|\mathbf{O}, m)]. \end{aligned} \tag{8}$$

By assuming that $p(m)$ is a uniform distribution, we obtain the proportion relation between $\widetilde{q}(m|\mathbf{O})$ and $\mathcal{F}^m$, and an optimal model structure in a sense of maximum a posterior probability can be selected as follows:

$$\widetilde{m} = \arg\max_{\{m\}} \widetilde{q}(m|\mathbf{O}) = \arg\max_{\{m\}} \mathcal{F}^m. \tag{9}$$

This indicates that by maximizing total $\mathcal{F}^m$ with respect to not only $q(\Theta|\mathbf{O}, m), q(Z|\mathbf{O}, m)$, but also m, we can obtain the optimal parameter distributions and can select the optimal model structure simultaneously [20], [21].

IV. Applying a VB approach to acoustic modeling

In this section, we apply the VB approach to a continuous density HMM (left-to-right HMM with a GMM for each state) [16], [17]. The continuous density HMM has been widely used to represent a phoneme category in acoustic models for speech recognition. We show concrete forms of the optimal VB posterior distributions for model parameters and the VB objective function. Since the formulations for the posterior distributions are common to all phoneme categories, we omit the category suffix c in this section to simplify the equation forms.

A. Output distribution $p(\mathbf{O}, \mathbf{S}, \mathbf{V}|\Theta, m)$ and prior distribution $p(\Theta|m)$

The output distribution of a continuous density HMM, which represents a phoneme acoustic model, is expressed by

$$p(\mathbf{O}, \mathbf{S}, \mathbf{V}|\Theta, m) = \prod_{t=1}^{T} a_{s^{t-1}s^t} w_{s^t v^t} b_{s^t v^t}(\mathbf{O}^t), \tag{10}$$

where $\mathbf{S}$ is a set of sequences of HMM states, $\mathbf{V}$ is a set of sequences of Gaussian mixture components, and s^t and v^t denote the state and mixture components at a frame t. Here, $\mathbf{S}$ and $\mathbf{V}$ are sets of discrete hidden variables, which are the concrete forms of $\mathbf{Z}$. The parameter a_{ij} denotes the state transition probability from state i to state j, and w_{jk} is the k-th weight factor of the Gaussian mixture for state j. In addition, $b_{jk}(\mathbf{O}^t)(= \mathcal{N}(\mathbf{O}^t|\boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}))$ denotes the Gaussian with the mean vector $\boldsymbol{\mu}_{jk}$ and covariance matrix $\boldsymbol{\Sigma}_{jk}$ defined as:

$$\begin{aligned} &\mathcal{N}(\mathbf{O}^t|\boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) \\ &\triangleq C_{\mathcal{N}} |\boldsymbol{\Sigma}_{jk}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{O}^t - \boldsymbol{\mu}_{jk})' \boldsymbol{\Sigma}_{jk}^{-1} (\boldsymbol{O}^t - \boldsymbol{\mu}_{jk})\right) \\ &= (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}_{jk}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{O}^t - \boldsymbol{\mu}_{jk})' \boldsymbol{\Sigma}_{jk}^{-1} (\boldsymbol{O}^t - \boldsymbol{\mu}_{jk})\right), \end{aligned} \tag{11}$$

where $|\cdot|$ and $'$ denote the determinant and the transpose of the matrix, respectively, while $\Theta = \{a_{ij}, w_{jk}, \boldsymbol{\mu}_{jk}, \Sigma_{jk}^{-1} | i, j = 1, ..., J, k = 1, ..., L\}$ is a set of output distribution parameters. Here, J denotes the number of states in an HMM sequence and L denotes the number of Gaussian components in a state.

Prior distribution is assumed to be a conjugate distribution and is expressed as follows:

$$\begin{aligned} & p(\Theta|m) \\ & = \prod_{i=1}^{J}\prod_{j=1}^{J}\prod_{k=1}^{L} p(\{a_{ij'}\}_{j'=1}^{J}|m)p(\{w_{jk'}\}_{k'=1}^{L}|m)p(b_{jk}|m) \\ & = \prod\nolimits_{i,j,k} \mathcal{D}(\{a_{ij'}\}_{j'=1}^{J}|\phi^0)\mathcal{D}(\{w_{jk'}\}_{k'=1}^{L}|\varphi^0) \\ & \qquad \mathcal{N}(\boldsymbol{\mu}_{jk}|\boldsymbol{\nu}_{jk}^0,(\xi^0)^{-1}\mathbf{\Sigma}_{jk})\prod_{d=1}^{D}\mathcal{G}(\mathbf{\Sigma}_{jk,d}^{-1}|\eta^0,\mathbf{R}_{jk,d}^0), \end{aligned} \tag{12}$$

where $b_{jk} = \{\boldsymbol{\mu}_{jk}, \mathbf{\Sigma}_{jk}^{-1}\}$. Here, $\Phi^0 = \{\phi^0, \varphi^0, \xi^0, \boldsymbol{\nu}^0, \eta^0, \mathbf{R}^0\}$ is a set of hyper-parameters, and is assumed to be a constant. In Eq. (12), $\mathcal{D}$ denotes a Dirichlet distribution and $\mathcal{G}$ denotes a gamma distribution. If covariance matrix elements are off the diagonal, a Wishart distribution is set as a prior distribution. The concrete forms of the distributions are defined as follows:

$$\begin{cases} \mathcal{D}(\{a_{ij}\}_{j=1}^{J}|\phi^0) \triangleq C_{\mathcal{D}}(\phi^0)\prod_{j}(a_{ij})^{\phi^0-1} \\ \mathcal{D}(\{w_{jk}\}_{k=1}^{L}|\varphi^0) \triangleq C_{\mathcal{D}}(\varphi^0)\prod_{k}(w_{jk})^{\varphi^0-1} \\ \mathcal{N}(\boldsymbol{\mu}_{jk}|\boldsymbol{\nu}_{jk}^0,(\xi^0)^{-1}\mathbf{\Sigma}_{jk}) \triangleq C_{\mathcal{N}}(\xi^0)|\mathbf{\Sigma}_{jk}|^{-\frac{1}{2}} \\ \qquad \exp\left(-\frac{\xi^0}{2}(\boldsymbol{\mu}_{jk}-\boldsymbol{\nu}_{jk}^0)'\mathbf{\Sigma}_{jk}^{-1}(\boldsymbol{\mu}_{jk}-\boldsymbol{\nu}_{jk}^0)\right) \\ \mathcal{G}(\mathbf{\Sigma}_{jk,d}^{-1}|\eta^0,\mathbf{R}_{jk,d}^0) \\ \qquad \triangleq C_{\mathcal{G}}(\eta^0,\mathbf{R}_{jk,d}^0)\left(\mathbf{\Sigma}_{jk,d}^{-1}\right)^{\frac{\eta^0}{2}-1}\exp\left(-\frac{\mathbf{R}_{jk,d}^0}{2\mathbf{\Sigma}_{jk,d}}\right) \end{cases} \tag{13}$$

where

$$\begin{cases} C_{\mathcal{D}}(\phi^0) \triangleq \Gamma(J\phi^0)/(\Gamma(\phi^0))^J \\ C_{\mathcal{D}}(\varphi^0) \triangleq \Gamma(L\varphi^0)/(\Gamma(\varphi^0))^L \\ C_{\mathcal{N}}(\xi^0) \triangleq (\xi^0/2\pi)^{\frac{D}{2}} \\ C_{\mathcal{G}}(\eta^0,\mathbf{R}_{jk,d}^0) \triangleq \left(\mathbf{R}_{jk,d}^0/2\right)^{\frac{\eta^0}{2}}/\Gamma(\eta^0/2) \end{cases} . \tag{14}$$

The setting of these output and prior distributions is the same as in [9].

B. Optimal VB posterior distribution for output distribution parameters $\widetilde{q}(\Theta|\mathbf{O},m)$

Since the prior distributions (Eq. (13)) are conjugate distributions, we can analytically obtain the optimal VB posterior distribution for output distribution parameters $\widetilde{q}(\Theta|\mathbf{O},m)$ (see [17] in detail), as follows:

$$\begin{aligned} & \widetilde{q}(\Theta|\mathbf{O},m) \\ & = \prod\nolimits_{i,j,k}\mathcal{D}(\{a_{ij'}\}_{j'=1}^{J}|\{\widetilde{\phi}_{ij'}\}_{j'=1}^{J})\mathcal{D}(\{w_{jk'}\}_{k'=1}^{L}|\{\widetilde{\varphi}_{jk'}\}_{k'=1}^{L}) \\ & \qquad \mathcal{N}(\boldsymbol{\mu}_{jk}|\widetilde{\boldsymbol{\nu}}_{jk},(\widetilde{\xi}_{jk})^{-1}\mathbf{\Sigma}_{jk})\prod\nolimits_{d}\mathcal{G}(\mathbf{\Sigma}_{jk,d}^{-1}|\widetilde{\eta}_{jk},\widetilde{\mathbf{R}}_{jk,d}). \end{aligned} \tag{15}$$

Note that Eqs. (12) and (15) are represented in the same function family, and the only difference is that the set of hyper-parameters Φ^0 in Eq. (12) is replaced with a set of posterior distribution parameters $\widetilde{\Phi} \equiv \{\widetilde{\phi}, \widetilde{\varphi}, \widetilde{\xi}, \widetilde{\boldsymbol{\nu}}, \widetilde{\eta}, \widetilde{\mathbf{R}}\}$ in Eq. (15). We adopt the conjugate prior distribution because the posterior distribution joins the same function family as the prior distribution theoretically and is obtained analytically, which is the characteristic of the exponential distribution family. Here, $\widetilde{\Phi}$ are defined as:

$$\begin{cases} \widetilde{\phi}_{ij} = \phi^0 + \widetilde{\gamma}_{ij}, \\ \widetilde{\varphi}_{jk} = \varphi^0 + \widetilde{\zeta}_{jk} \\ \widetilde{\xi}_{jk} = \xi^0 + \widetilde{\zeta}_{jk} \\ \widetilde{\boldsymbol{\nu}}_{jk} = \left(\xi^0\boldsymbol{\nu}_{jk}^0 + \sum_{t=1}^{T}\widetilde{\zeta}_{jk}^t\mathbf{O}^t\right)/\widetilde{\xi}_{jk} \\ \widetilde{\eta}_{jk} = \eta^0 + \widetilde{\zeta}_{jk} \\ \widetilde{\mathbf{R}}_{jk,d} = \mathbf{R}_{jk,d}^0 + \xi^0(\boldsymbol{\nu}_{jk,d}^0 - \widetilde{\boldsymbol{\nu}}_{jk,d})^2 + \sum_{t=1}^{T}\widetilde{\zeta}_{jk}^t(\mathbf{O}_d^t - \widetilde{\boldsymbol{\nu}}_{jk,d})^2 \end{cases} \tag{16}$$

where $\widetilde{\gamma}_{ij}$, $\widetilde{\zeta}_{e,jk}^t$, and $\widetilde{\zeta}_{jk}$ are the sufficient statistics of a continuous density HMM, and defined as follows:

$$\begin{cases} \widetilde{\gamma}_{ij}^t \triangleq \widetilde{q}(s^{t-1}=i,s^t=j|\mathbf{O},m) \\ \widetilde{\gamma}_{ij} \triangleq \sum_{t=1}^{T}\widetilde{\gamma}_{ij}^t \\ \widetilde{\zeta}_{jk}^t \triangleq \widetilde{q}(s^t=j,v^t=k|\mathbf{O},m) \\ \widetilde{\zeta}_{jk} \triangleq \sum_{t=1}^{T}\widetilde{\zeta}_{jk}^t \end{cases} . \tag{17}$$

Therefore, $\widetilde{\Phi}$ can be calculated from Φ^0, $\widetilde{\gamma}_{e,ij}^t$ and $\widetilde{\zeta}_{e,jk}^t$, enabling $\widetilde{q}(\Theta|\mathbf{O},m)$ to be obtained.

C. Optimal VB posterior distribution for hidden variables $\widetilde{q}(\mathbf{S},\mathbf{V}|\mathbf{O},m)$

From the output distribution and prior distribution in Section IV-A, the optimal VB posterior distribution for hidden variables $\widetilde{q}(\mathbf{S},\mathbf{V}|\mathbf{O},m)$ is represented as follows:

$$\begin{aligned} & \widetilde{q}(\mathbf{S},\mathbf{V}|\mathbf{O},m) \\ & \propto \prod_{t}\exp\left(\langle\log a_{s^{t-1}s^t}\rangle_{\widetilde{q}(\{a_{ij'}\}_{j'=1}^{J}|\boldsymbol{O},m)}\right) \\ & \qquad \exp\left(\langle\log w_{s^tv^t}\rangle_{\widetilde{q}(\{w_{jk'}\}_{k'=1}^{L}|\boldsymbol{O},m)}\right) \\ & \qquad \exp\left(\langle\log b_{s^tv^t}(\boldsymbol{O}^t)\rangle_{\widetilde{q}(b_{jk}|\boldsymbol{O},m)}\right). \end{aligned}$$

Therefore, the optimal VB posterior distribution for hidden variables $\widetilde{q}(\mathbf{S},\mathbf{V}|\mathbf{O},m)$ is obtained by considering a normalization constant as follows:

$$\widetilde{q}(\mathbf{S},\mathbf{V}|\mathbf{O},m) = \frac{\prod_{t=1}^{T}\widetilde{a}_{s^{t-1}s^t}\widetilde{w}_{s^tv^t}\widetilde{b}_{s^tv^t}(\mathbf{O}^t)}{\sum_{\mathbf{S},\mathbf{V}}\prod_{t=1}^{T}\widetilde{a}_{s^{t-1}s^t}\widetilde{w}_{s^tv^t}\widetilde{b}_{s^tv^t}(\mathbf{O}^t)} \tag{18}$$

where

$$\begin{aligned}
&\widetilde{a}_{s^{t-1}s^t} = \exp\Big(\Psi(\widetilde{\phi}_{s^{t-1}s^t}) - \Psi(\textstyle\sum_{s^{t'}} \widetilde{\phi}_{s^{t-1}s^{t'}})\Big),\\
&\widetilde{w}_{s^t v^t} = \exp\Big(\Psi(\widetilde{\varphi}_{s^t v^t}) - \Psi(\textstyle\sum_{v^{t'}} \widetilde{\varphi}_{s^t v^{t'}})\Big),\\
&\widetilde{b}_{s^t v^t}(\mathbf{O}^t) = \exp\Bigg(\frac{D}{2}\Bigg(-\log 2\pi - \frac{1}{\widetilde{\xi}_{s^t v^t}} + \Psi(\frac{\widetilde{\eta}_{s^t v^t}}{2})\Bigg)\\
&\qquad - \frac{1}{2}\sum_d \Bigg(\log\Bigg(\frac{\widetilde{\mathbf{R}}_{s^t v^t,d}}{2}\Bigg) + \frac{(\mathbf{O}_d^t - \widetilde{\boldsymbol{\nu}}_{s^t v^t,d})^2 \widetilde{\eta}_{s^t v^t}}{\widetilde{\mathbf{R}}_{s^t v^t,d}}\Bigg)\Bigg).
\end{aligned} \tag{19}$$

where $\Psi(y)$ is a digamma function defined as $\Psi(y) \equiv \partial/\partial y \log \Gamma(y)$.

From $\widetilde{a}_{ij}, \widetilde{w}_{jk}$ and $\widetilde{b}_{jk}(\mathbf{O}_e^t)$, we can obtain the transition and occupation probabilities $\widetilde{\gamma}_{ij}^t$ and $\widetilde{\zeta}_{jk}^t$, which are required for the calculation of $\widetilde{q}(\Theta|\mathbf{O}, m)$ in Section IV-B. From the variational calculation, $\widetilde{\gamma}_{ij}^t$ is obtained as follows:

$$\widetilde{\gamma}_{ij}^t = \frac{\widetilde{\alpha}_i^{t-1}\widetilde{a}_{ij} \sum_k \widetilde{w}_{jk}\widetilde{b}_{jk}(\mathbf{O}^t)\widetilde{\beta}_j^t}{\sum_{j'} \widetilde{\alpha}_{j'}^T}, \tag{20}$$

where $\widetilde{\alpha}$ and $\widetilde{\beta}$ are VB forward and backward probabilities defined as:

$$\left\{\begin{array}{lll} \widetilde{\alpha}_j^t & \equiv & \left(\sum_i \widetilde{\alpha}_i^{t-1}\widetilde{a}_{ij}\right) \sum_k \widetilde{w}_{jk}\widetilde{b}_{jk}(\mathbf{O}^t) \\ \widetilde{\beta}_j^t & \equiv & \sum_i \widetilde{a}_{ji}\left(\sum_k \widetilde{w}_{ik}\widetilde{b}_{ik}(\mathbf{O}^{t+1})\right)\widetilde{\beta}_i^{t+1} \end{array}\right. . \tag{21}$$

$\widetilde{\alpha}_j^{t=0}$ and $\widetilde{\beta}_j^{t=T}$ are initialized appropriately. Similarly, $\widetilde{\zeta}_{jk}^t$ is obtained from the variational calculation as follows:

$$\widetilde{\zeta}_{jk}^t = \frac{\left(\sum_i \widetilde{\alpha}_i^{t-1}\widetilde{a}_{ij}\right) \widetilde{w}_{jk}\widetilde{b}_{jk}(\mathbf{O}^t)\widetilde{\beta}_j^t}{\sum_i \widetilde{\alpha}_i^T}. \tag{22}$$

Thus, $\widetilde{\gamma}_{ij}^t$ and $\widetilde{\zeta}_{jk}^t$ are calculated efficiently by using a probabilistic assignment via the familiar *forward-backward algorithm*.

Analogous to the forward-backward algorithm, the *Viterbi algorithm* is also derived by exchanging a summation over i for maximization over i in the calculation of VB forward probability $\widetilde{\alpha}_j^t$ in Eq. (21).

D. VB objective function $\mathcal{F}^m$

In this section, we discuss VB objective function $\mathcal{F}^m$, which is a criterion for both posterior distribution estimation and model structure optimization, and provide general calculation results. By substituting the VB posterior distribution obtained in Sections IV-B and IV-C, we obtain analytical results for $\mathcal{F}^m$. Although we focus on one phoneme category in this section, the total $\mathcal{F}^m$ for all categories is obtained by simply summing up the $\mathcal{F}^m$ results obtained in this section for all categories, according to Eq. (7). We can separate $\mathcal{F}^m$ into two components: one is composed of only $\widetilde{q}(S, V|\mathbf{O}, m)$, whereas the other is mainly composed of $\widetilde{q}(\Theta|\mathbf{O}, m)$. Therefore, we define $\mathcal{F}_\Theta^m$ and $\mathcal{F}_{\mathbf{S},\mathbf{V}}^m$, and represent $\mathcal{F}^m$ as follows:

$$\begin{aligned}
\mathcal{F}^m &= -\sum_{\mathbf{S},\mathbf{V}} \widetilde{q}(\mathbf{S}, \mathbf{V}|\mathbf{O}, m) \log \widetilde{q}(\mathbf{S}, \mathbf{V}|\mathbf{O}, m)\\
&\quad + \left\langle \sum_{\mathbf{S},\mathbf{V}} \widetilde{q}(\mathbf{S}, \mathbf{V}|\mathbf{O}, m) \log \frac{p(\mathbf{O}, \mathbf{S}, \mathbf{V}|\Theta, m)p(\Theta|m)}{\widetilde{q}(\Theta|\mathbf{O}, m)} \right\rangle_{\widetilde{q}(\Theta|\boldsymbol{O},m)}\\
&\triangleq -\mathcal{F}_{\mathbf{S},\mathbf{V}}^m + \mathcal{F}_\Theta^m.
\end{aligned} \tag{23}$$

$\mathcal{F}_{\mathbf{S},\mathbf{V}}^m$ is an entropy value and is calculated at the E-step in the VB EM algorithm. $\mathcal{F}_\Theta^m$ is obtained as follows:

$$\begin{aligned}
\mathcal{F}_\Theta^m &= \sum_i \log \frac{\Gamma\left(J\phi^0\right)\prod_{j'}\Gamma(\widetilde{\phi}_{ij'})}{\Gamma\left(\sum_{j'}\widetilde{\phi}_{ij'}\right)\Gamma(\phi^0)^J}\\
&\quad + \sum_j \log \frac{\Gamma\left(L\varphi^0\right)\prod_{k'}\Gamma(\widetilde{\varphi}_{jk'})}{\Gamma\left(\sum_{k'}\widetilde{\varphi}_{jk'}\right)\Gamma(\varphi^0)^L}\\
&\quad + \sum_{j,k} \log \left\{ (2\pi)^{-\frac{\widetilde{\zeta}_{jk}D}{2}} \left(\frac{\xi^0}{\widetilde{\xi}_{jk}}\right)^{\frac{D}{2}} \frac{2^{\frac{\widetilde{\eta}_{jk}D}{2}} \left(\Gamma\left(\frac{\widetilde{\eta}_{jk}}{2}\right)\right)^D \left|\mathbf{R}_{jk}^0\right|^{\frac{\eta^0}{2}}}{2^{\frac{\eta^0 D}{2}} \left(\Gamma\left(\frac{\eta^0}{2}\right)\right)^D \left|\widetilde{\mathbf{R}}_{jk}\right|^{\frac{\widetilde{\eta}_{jk}}{2}}} \right\}.
\end{aligned} \tag{24}$$

This objective function is used as an optimization criterion with respect to model structure m.

E. Bayesian predictive classification using VB posterior distributions

After acoustic modeling, we obtain the optimal VB posterior distributions for the optimal model structure $\widetilde{q}(\Theta|\mathbf{O}, \widetilde{m})$. In recognition, an input speech $\mathbf{x}^t$ for a frame t is classified as the optimal phoneme class $\bar{c}$ using $p(c|\mathbf{x}^t, \mathbf{O}, \widetilde{m})$ for the estimated model structure $\widetilde{m}$ defined as follows:

$$\bar{c} = \arg\max_{\{c\}} p(c|\mathbf{x}^t, \mathbf{O}, \widetilde{m}) \equiv \arg\max_{\{c\}} p(c)p(\mathbf{x}^t|c, \mathbf{O}, \widetilde{m}). \tag{25}$$

Here, $p(c)$ is the class prior distribution obtained by language and lexicon models, and $p(\mathbf{x}^t|c, \mathbf{O}, \widetilde{m})$ is the predictive posterior distribution. When we approximate the true posterior distribution $p(\Theta|c(j), \mathbf{O}, \widetilde{m})$ by using the estimated VB posterior distributions $\widetilde{q}(\Theta|c(j), \mathbf{O}, \widetilde{m})$, $p(\mathbf{x}^t|c(j), \mathbf{O}, \widetilde{m})$ can be approximated as

$$p(\mathbf{x}^t|c(j), \mathbf{O}, \widetilde{m}) \approx \int d\Theta p(\mathbf{x}^t|c(j), \Theta, \widetilde{m})\widetilde{q}(\Theta|c(j), \mathbf{O}, \widetilde{m}). \tag{26}$$

We focus on the integral part. The integral over w_{jk}, $\boldsymbol{\mu}_{jk}$ and Σ_{jk}^{-1} for a frame can be solved analytically and found to be a mixture Student-t distribution, as follows:

$$\begin{aligned}
&\int p(\mathbf{x}^t|c(j), \Theta, \widetilde{m})\widetilde{q}(\Theta|c(j), \mathbf{O}, \widetilde{m})d\Theta\\
&= \sum_k \frac{\widetilde{\varphi}_{jk}}{\sum_{k'}\widetilde{\varphi}_{jk'}} \prod_d \mathcal{T}(\mathbf{x}_d^t|\widetilde{\boldsymbol{\nu}}_{jk,d}, (1+\widetilde{\xi}_{jk})\widetilde{\mathbf{R}}_{jk,d}\widetilde{\eta}_{jk}/\widetilde{\xi}_{jk}, \widetilde{\eta}_{jk}).
\end{aligned} \tag{27}$$

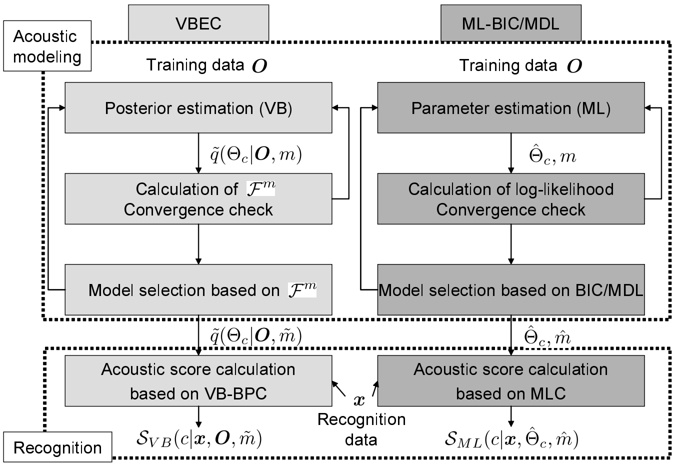

VBEC
ML-BIC/MDL
Acoustic modeling
Training data O
Training data O
Posterior estimation (VB)
Parameter estimation (ML)
$\tilde{q}(\Theta_c|\boldsymbol{O},m)$
$\hat{\Theta}_c, m$
Calculation of $\mathcal{F}^m$ Convergence check
Calculation of log-likelihood Convergence check
Model selection based on $\mathcal{F}^m$
Model selection based on BIC/MDL
$\tilde{q}(\Theta_c|\boldsymbol{O},\tilde{m})$
$\hat{\Theta}_c, \hat{m}$
Acoustic score calculation based on VB-BPC
Acoustic score calculation based on MLC
$\boldsymbol{x}$
Recognition data
$\mathcal{S}_{VB}(c|\boldsymbol{x},\boldsymbol{O},\tilde{m})$
$\mathcal{S}_{ML}(c|\boldsymbol{x},\hat{\Theta}_c,\hat{m})$
Recognition

Fig. 1. Total speech recognition frameworks based on VBEC and ML-BIC/MDL. $\mathcal{S}_{VB}(c|\mathbf{x},\mathbf{O},\widetilde{m})$ represents the VB-BPC score and $\mathcal{S}_{ML}(c|\mathbf{x},\widehat{\Theta}_c,\widehat{m})$ represents the Maximum Likelihood based Classification (MLC) score of a phoneme category c for recognition data $\mathbf{x}$.

The index c is removed in Eq. (27) to avoid a complicated equation. Student-t distribution is defined as follows:

$$\begin{aligned}&\mathcal{T}(\mathbf{x}_d^t|\widetilde{\boldsymbol{\nu}}_{jk,d},(1+\widetilde{\xi}_{jk})\widetilde{\mathbf{R}}_{jk,d}\widetilde{\eta}_{jk}/\widetilde{\xi}_{jk},\widetilde{\eta}_{jk})\\ &\triangleq \frac{\Gamma((\widetilde{\eta}_{jk}+1)/2)}{\Gamma(\widetilde{\eta}_{jk}/2)\Gamma(1/2)}\left(\frac{\widetilde{\xi}_{jk}}{(1+\widetilde{\xi}_{jk})\widetilde{\mathbf{R}}_{jk,d}}\right)^{1/2} \qquad (28)\\ &\quad\left(1+\frac{\widetilde{\xi}_{jk}}{(1+\widetilde{\xi}_{jk})\widetilde{\mathbf{R}}_{jk,d}}(\mathbf{x}_d^t-\widetilde{\boldsymbol{\nu}}_{jk,d})^2\right)^{-(\widetilde{\eta}_{jk}+1)/2}.\end{aligned}$$

Therefore, input speech can be classified by using the predictive score obtained from Eq. (27). We call this approach Bayesian Predictive Classification using VB posterior distributions (VB-BPC).

VB-BPC accomplishes the VBEC to be a fully 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. Figure 1 shows the VBEC framework compared with a conventional approach, ML-BIC/MDL: the model parameter estimation, model selection and speech classification are based on ML, BIC/MDL and Maximum Likelihood based Classification (MLC), respectively. The VBEC mitigates the “over-training 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.

V. Experiments

We conducted experiments to prove the effectiveness of VBEC and other Bayesian approaches in effective utilization of prior knowledge, automatic determination of acoustic model topologies, and marginalization effect. All the experiments in this paper were performed using the SOLON speech recognition toolkit [22] developed by NTT Communication Science Laboratories.

TABLE II
Training and test data and language model for JNAS

Training data	JNAS 20,000 utterances, 34 hours (male)
Test data	JNAS 100 utterances, 1,583 words (male)
Language model	Standard trigram (10 years of newspapers)
Vocabulary size	20,000
Perplexity	64.0

TABLE III
Experimental conditions

Sampling rate	16 kHz (16-bit quantization)
Feature vector (26 dimensions)	12 - order MFCC + Δ MFCC + Energy + Δ Energy
Window	Hamming
Frame size/shift	25/10 ms
Number of HMM states	3 (Left to right)
Number of phoneme categories	43

A. Effective utilization of prior knowledge

We employed Japanese Newspaper Article Sentences (JNAS) for the experiment. The quantitative features of the training and test sets are summarized in Tables II. Other experimental conditions are summarized in Table III.

The left side of Figure 2 shows a comparison of VBEC and ML-BIC/MDL with varying amounts of data and an enlarged view for more than 1,000 utterances is shown in the right side. VBEC performed as well as or better than ML-BIC/MDL with every amount of data. In particular, VBEC significantly outperformed the ML-BIC/MDL approaches for various tuning parameters (from $\lambda = 1$ to $\lambda = 4$) when the amounts of training data were small. Consequently, VBEC exhibited considerable superiority especially with small amounts of training data (less than 1,000 utterances), which solves the over-training problem.

B. Automatic determination of acoustic model topologies

Based on the model selection function by using the VB objective function, we can perform VBEC-based efficient model search algorithm and GMM-based decision tree clustering utilizing the acoustic model characteristics [23]. In these experiments, the availability of VBEC automatic determination was examined experimentally using various speech data. This

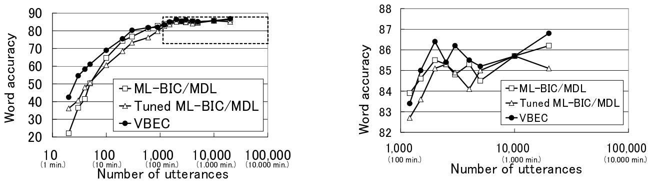


Fig. 2. The left side figure shows recognition rates according to the amounts of training data based on VBEC, ML-BIC/MDL and tuned ML-BIC/MDL. The right side figure shows an enlarged view of the left side figure for more than 1,000 utterances The horizontal axis is scaled logarithmically.

experimental section provides three subsections to confirm the robustness with respect to different speaking styles and languages, which are representative of speech variations, and of different test data by recognizing question utterances for a question answering system using an acoustic model trained by the Japanese read speech data set, whose conditions are mismatched with those of test data set.

1) Speaking style variation: First, we focused on speaking style variation of the training data set by preparing an isolated word speech (100 city names provided by JEIDA), LVCSR based on Japanese read speech (JNAS: Japanese Newspaper Article Sentences) and LVCSR based on Japanese lecture speech (CSJ: Corpus of Spontaneous Japanese). Speaking style greatly influences acoustic features, and acoustic models need to be constructed *manually* depending on the style when the ML method is used. However, VBEC determination could allow us to replace manual construction with automatic construction for various speaking styles. Therefore, we examined the robustness of VBEC determination for various speaking styles. The configuration of feature extraction was 12-order MFCC + Δ MFCC (24 dim.) for 100 city names, 12-order MFCC + Δ MFCC + Energy + Δ Energy (26 dim.) for JNAS and 12-order MFCC + Δ MFCC + Δ Energy (25 dim.) + CMN for CSJ. The sampling rate was 16 kHz, the frame size was 25 ms and the frame shift was 10 ms. For JNAS and CSJ, we used standard trigram models with vocabularies of 20,000 and 30,000, respectively. For the 100 city name task, the training data consisted of about 3,000 Japanese sentences (4.1 hours) spoken by 30 males and the recognition data consisted of 100 Japanese city names spoken by 25 males (a total of 2,400 words). For the JNAS task, the training data consisted of about 20,000 Japanese sentences (34 hours) spoken by 122 males and the recognition data consisted of 100 Japanese sentences spoken by 10 males (a total of about 2,000 words). For the CSJ task, the training data consisted of about 800 Japanese lectures (190 hours) spoken by 200 males and the recognition data consisted of 10 Japanese lectures spoken by 10 males (a total of about 27,000 words).

First, we examined the recognition performance of conventional ML-based acoustic models with manually varied model topologies for a number of clustered states and GMM components per state, which we use as baselines with which to compare the performance of the automatically determined model topology. The contour maps in Figures 3 and 4, and the white bar in Figure 5 show the recognition performance obtained with the ML method. Then, we provided the model, whose topologies were determined by VBEC, with recognition performance. For all the tasks, the resultant combinations of the numbers of states and components per state, determined by VBEC, were included in the high performance area in Figures 3, 4 and 5. In addition, the recognition performance (97.9 %, 91.7 WACC and 74.5 WACC) of all the tasks reached the highest performance (98.0 %, 91.4 WACC and 74.2 WACC) obtained with ML methods. Consequently, we confirmed that VBEC determination is effective for various speaking styles, namely isolated word speech, continuous read speech and spontaneous lecture speech.

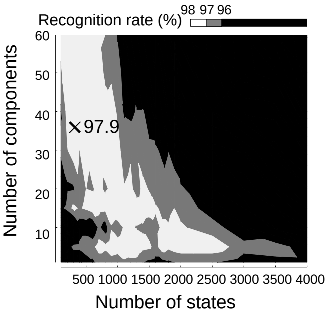


Fig. 3. 100 city name.

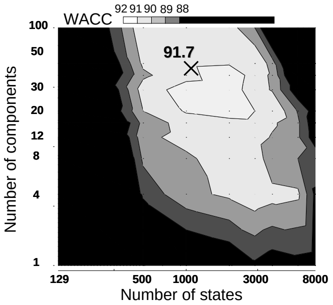
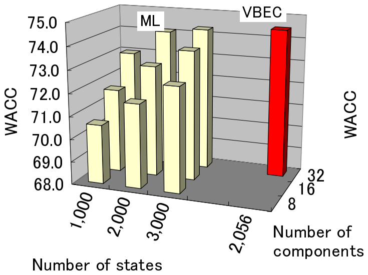


Fig. 4. JNAS.

Determined model topologies and their word accuracies



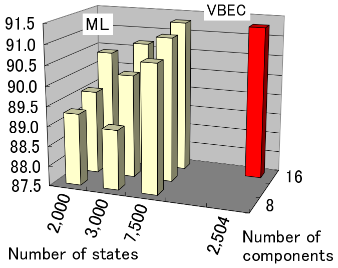


Fig. 5. CSJ.

Fig. 6. WSJ.

Determined model topologies and their word accuracies

2) Language variation: In our second set of experiments, we focused on the effect on VBEC determination of language variation. The acoustic feature depends strongly on the languages, and the appropriate model topology will be changed depending on the language. Therefore, we must examine how VBEC determination works even for a different language task. We used English read speech (WSJ: Wall Street Journal) as a different language task from Japanese tasks. The feature extraction configuration is 12-order MFCC + Δ MFCC + ΔΔ MFCC + Energy + Δ Energy + ΔΔ Energy (39 dim.) + CMN. The other configuration was the same as that in Section V-B1. We used a standard trigram model that had a vocabulary of 20,000. The training data consisted of about 20,000 English sentences (36 hours) spoken by 143 males and the recognition data consisted of 100 English sentences spoken by 5 males (a total of about 2,000 words).

As in Section V-B1, we prepared the recognition performance of conventional ML-based acoustic models with manually varied model topologies for a number of clustered states and GMM components per state. The white bar in Figure 6 shows the recognition performance obtained by the ML method and the black bar represents the VBEC determined model with the recognition performance. Although the determined model topology with 2,504 states and 32 components was far from the best ML results of 7,000 states and 32 components, its performance (91.3 WACC) matched the best ML performance (91.3 WACC), and we can say that VBEC determination is effective even for a different language task such as English rather than Japanese. In addition, the VBEC determined model exhibited the best ML performance with less than half the total number of Gaussians, which reduced the decoding time to less than half (8.29 RTF → 2.35RTF).

C. Marginalization effect

We examine the effectiveness of VB-BPC for supervised speaker adaptation as a practical application of the VB-BPC, which shows its superiority for solving the sparse data problem. We compare the improvement in the adaptation accuracy using VB-BPC, VB-BPC-MEAN, UBPC (Uniform posterior based BPC [14]) and δBPC(corresponding to MAP adaptation [9]), each of which belongs to the direct HMM parameter adaptation scheme. Table IV summarizes the experimental conditions. The initial (prior) acoustic model was constructed by read sentences and we adapted this model using 10 lectures spoken by 10 males and their labels [24]. In this task, the mismatch between training and adaptation data is caused not only by the speakers, but also by the difference in speaking styles between read speech and a lecture. The total training data for the initial models consisted of 10,709 Japanese utterances spoken by 44 males. In the initial model training, we constructed a speaker-independent model based on 1,000 context-dependent HMM states, using a phonetic decision tree method. The output distribution in each state was represented by a 16-component mixture distribution, and the model parameters were trained based on conventional ML estimation. Each lecture was divided in half based on the utterance units, and the first half of the lecture was used as adaptation data and the second half was used as recognition data. The total adaptation data consisted of more than 60 utterances for each male, and 1, 2, 4, 8, 16, 32, 40, 48 and 60 utterances were used as adaptation data. As a result, about 9 sets of adapted acoustic models for several amounts of adaptation data were prepared for each male. The prior parameter settings are shown in Table V, and were used to estimate the MAP parameters in δBPC(MAP) and UBPC, and also to estimate the VB posteriors in VB-BPC-MEAN and VB-BPC. When setting the UBPC hyper-parameters, we optimized the hyper-parameters in advance by trying eight kinds of combinations of $C = 2, 3, 4$ and 5 and $\rho = 0.7$ and 0.9 with reference to the result in [14], and adopted a combination of $\{C = 3, \rho = 0.9\}$, which provided the best average word accuracy. Throughout this experiment we used a beam search algorithm with sufficient beam width and a sufficient number of hypotheses to avoid search errors in decoding. The language model weight used in this experiment was optimized by the word accuracy of each result.

Figure 7 compares the recognition results obtained with VB-BPC, VB-BPC-MEAN, UBPC and δBPC(MAP) for several amounts of adaptation data with the baseline performance for the non-adapted speaker independent model (62.9 percent word accuracy). First, we focus on the effectiveness of the marginalization of the model parameters in BPCs for the sparse data problem. Namely, we compared the results of VB-BPC, VB-BPC-MEAN and UBPC with that of δBPC(MAP), which does not marginalize the model parameters at all. From Figure 7, we found that for a small amount of adaptation data (fewer than 8 adaptation utterances), VB-BPC, VB-BPC-MEAN and UBPC were better than δBPC(MAP), which confirms the effectiveness of the marginalization of the model parameters. By examining the results in this region in further detail, VB-BPC was better than UBPC by 0.7 $\sim$ 1.5 points, and VB-BPC-MEAN and UBPC behaved similarly. This suggests the effectiveness of the wide tail property of the Student's t-distribution, which is obtained by the marginalization of the variance parameters in addition to the mean parameters. Second, for any given amount of adaptation data, VB-BPC and VB-BPC-MEAN achieved comparable or better performance than UBPC, which required hyper-parameter (C and ρ) optimization. Therefore, we can say that VB-BPC and VB-BPC-MEAN could determine the shapes of their distributions automatically and appropriately from the adaptation data without tuning the hyper-parameters. Finally, VB-BPC was the best for almost all amounts of adaptation data. VB-BPC approached the δBPC(MAP) performance asymptotically, and provided the highest score of 72.9 percent word accuracy for this task (the benchmark score obtained by the speaker independent acoustic

TABLE IV
EXPERIMENTAL CONDITIONS FOR SPEAKER ADAPTATION

Sampling rate/quantization	16 kHz / 16 bit
Feature vector	12 order MFCC with energy
(39 dimensions)	$+\Delta+\Delta\Delta$
Window	Hamming
Frame size/shift	25/10 ms
Num. of states	3 (Left to right)
Num. of phoneme categories	43
Num. of phonetic questions	144
Num. of mixture components	16

Initial training data	ASJ: 10,709 utterances, 10.2 hours (44 males)
Adaptation data	CSJ: 1st-half lectures (10 males)
Test data	CSJ: 2nd-half lectures (10 males)

Language model	Standard trigram (made by CSJ transcription)
Vocabulary size	30, 000
Perplexity	82.2
OOV rate	2.1 %

CSJ: Corpus of Spontaneous Japanese

TABLE V
PRIOR PARAMETER SETTING

φ_{jk}^{0}	10
ξ_{jk}^{0}	10
$\boldsymbol{\nu}_{jk}^{0}$	SI mean vector of Gaussian k in state j
η_{jk}^{0}	10
R_{jk}^{0}	SI covariance matrix of Gaussian k in state j $\times \eta_{jk}^{0}$

SI: Speaker Independent

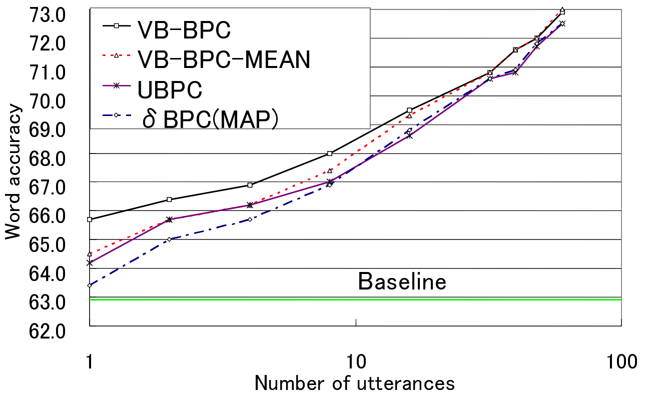


Fig. 7. Word accuracy for various amounts of adaptation data. The horizontal axis is scaled logarithmically.

model trained by using lectures is about 72.0 percent word accuracy in [24]). This confirms the steady improvement in the performance obtained using VB-BPC.

Thus, we show the effectiveness of marginalization using the VB-BPC-based Student's t-distribution for the sparse data problem.

VI. Summary and related work

This paper introduces applications of Bayesian approaches to speech recognition, especially for Variational Bayesian Estimation and Clustering for speech recognition (VBEC). The experiments proved the effectiveness of VBEC compared to the other Bayesian approaches for efficient utilization of prior knowledge, automatic determination of acoustic model topologies, and marginalization effect.

Currently, VB becomes a common technique in speech processing. Table VI summarizes the technical trend in VB-applied speech information processing. Note that VB has been widely applied to speech recognition and other forms of speech processing. Given such a trend, VBEC plays an important role in pioneering the main formulation and implementation of VB based speech recognition, which is a core technology in this field. In addition, other Bayesian approaches than VB are effectively applied to speech recognition, e.g., on-line Bayesian adaptation [25], [26], structural Bayes [27], [28], quasi-Bayes [29]–[31], and evidence framework [32]. The variational techniques used in VB are also applied to some speech recognition approaches [33], [34]. These approaches are associated with the progress of Bayesian approaches in statistics and machine learning fields, and speech recognition based on Bayesian approaches will advance further using the recent progress in these fields (e.g., Markov chain Monte Carlo, non-parametric Bayes).

VII. Acknowledgements

I thank Seong-Jun Hahm, Marc Delcroix, and Roland Roller at NTT Communication Science Laboratories for fruitful discussion.

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society B*, vol. 39, pp. 1–38, 1976.

[2] F. Jelinek, "Continuous speech recognition by statistical methods," in *Proc. IEEE*, 1976, vol. 64(4), pp. 532–556.

[3] X. D. Huang, Y. Ariki, and M. A. Jack, *Hidden Markov models for speech recognition*, Edinburgh University Press, 1990.

[4] X. D. Huang, A. Acero, and H. W. Hon, *Spoken language processing, a guide to theory, algorithm, and system development*, Prentice Hall PTR, 2001.

[5] S. Furui, "Recent advances in spontaneous speech recognition and understanding," in *Proc. SSPR2003*, 2003, pp. 1–6.

[6] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis, Second Edition*, Springer-Verlag, 1985.

[7] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*, John Wiley & Sons Ltd, 1994.

[8] C.-H. Lee, C. H. Lin, and B-H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 39, pp. 806–814, 1991.

[9] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[10] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *Journal of the Acoustical Society of Japan (E)*, vol. 21, pp. 79–86, 2000.

[11] W. Chou and W. Reichl, "Decision tree state tying based on penalized Bayesian information criterion," in *Proc. ICASSP1999*, 1999, vol. 1, pp. 345–348.

[12] S. Chen and R. Gopinath, "Model selection in acoustic modeling," in *Proc. Eurospeech1999*, 1999, vol. 3, pp. 1087–1090.

[13] K. Shinoda and K. Iso, "Efficient reduction of Gaussian components using MDL criterion for HMM-based speech recognition," in *Proc. ICASSP2001*, 2001, vol. 1, pp. 869–872.

[14] H. Jiang, K. Hirose, and Q. Huo, "Robust speech recognition based on a Bayesian prediction approach," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 426–440, 1999.

[15] Q. Huo and C.-H. Lee, "A Bayesian predictive classification approach to robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 200–204, 2000.

[16] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, *Application of variational Bayesian approach to speech recognition*, NIPS 2002, MIT Press, 2002.

[17] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Variational Bayesian estimation and clustering for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 365–381, 2004.

TABLE VI
Technical trend of speech recognition using variational Bayes

Topic	References
Feature extraction	[35], [36]
Voice activity detection	[37]
Speech GMM for noise robust ASR	[38]
Clustering context-dependent HMM states	[39]–[43]
Formulation of Bayesian speech recognition	[16], [17], [32], [44]
Selection of number of GMM components	[45]–[47]
Acoustic model adaptation	[48]–[52]
Determination of acoustic model topology	[23], [53]
Non-parametric Bayes for acoustic models	[54], [55]
Gaussian reduction in acoustic models	[56]
Bayesian prediction for speech recognition	[51], [57]
Language modeling	[58]–[60]
Statistical speech synthesis	[61], [62]

[18] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, pp. 183–233, 1997.

[19] S. Waterhouse, D. MacKay, and T. Robinson, *Bayesian methods for mixtures of experts*, NIPS 7, MIT Press, 1995.

[20] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," in *Proc. Uncertainty in Artificial Intelligence (UAI) 15*, 1999.

[21] N. Ueda and Z. Ghahramani, "Bayesian model search for mixture models based on optimizing variational bounds," *Neural Networks*, vol. 15, pp. 1223–1241, 2002.

[22] T. Hori, "NTT Speech recognizer with OutLook On the Next generation: SOLON," in *Proc. NTT Workshop on Communication Scene Analysis*, 2004, vol. 1, SP-6.

[23] S. Watanabe, A. Sako, and A. Nakamura, "Automatic determination of acoustic model topology using variational Bayesian estimation and clustering for large vocabulary continuous speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, 2006, 855-872.

[24] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui, "Benchmark test for speech recognition using the Corpus of Spontaneous Japanese," in *Proc. SSPR2003*, 2003, pp. 135–138.

[25] Q. Huo and C.-H. Lee, "On-line adaptive learning of the correlated continuous density hidden Markov models for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 386–397, 1998.

[26] S. Watanabe and A. Nakamura, "Predictor–corrector adaptation by using time evolution system with macroscopic time scale," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 395–406, 2010.

[27] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 276–287, 2001.

[28] O. Siohan, T.A. Myrvoll, and C.H. Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation," *Computer Speech & Language*, vol. 16, no. 1, pp. 5–24, 2002.

[29] U. E. Makov and A. F. M. Smith, "A quasi-Bayes unsupervised learning procedure for priors," *IEEE Transactions on Information Theory*, vol. 23, pp. 761–764, 1977.

[30] Q. Huo, C. Chan, and C.-H. Lee, "On-line adaptation of the SCHMM parameters based on the segmental quasi-Bayes learning for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 141–144, 1996.

[31] J. T. Chien, "Quasi-Bayes linear regression for sequential learning of hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 268–278, 2002.

[32] Y. Zhang, P. Liu, J.T. Chien, and F. Soong, "An evidence framework for bayesian learning of continuous-density hidden markov models," in *Proc. ICASSP 2009*, 2009, pp. 3857–3860.

[33] J.R. Hershey and P.A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *Proc. ICASSP 2007*, 2007, pp. 317–320.

[34] Y. Kubo, S. Watanabe, A. Nakamura, and T. Kobayashi, "A regularized discriminative training method of acoustic models derived by minimum relative entropy discrimination," in *Proc. Interspeech 2010*, 2010, pp. 2954–2957.

[35] O. Kwon, T.-W. Lee, and K. Chan, "Application of variational Bayesian PCA for speech feature extraction," in *Proc. ICASSP2002*, 2002, vol. 1, pp. 825–828.

[36] F. Valente and C. Wellekens, "Variational Bayesian feature selection for Gaussian mixture models," in *Proc. ICASSP2004*, 2004, vol. 1, pp. 513–516.

[37] D. Cournapeau, S. Watanabe, A. Nakamura, and T. Kawahara, "Online unsupervised classification with model comparison in the variational bayes framework for voice activity detection," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 6, pp. 1071–1083, 2010.

[38] S.G.S. Pettersen, *Robust Speech Recognition in the Presence of Additive Noise*, Ph.D. thesis, Norwegian University of Science and Technology, 2008.

[39] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Constructing shared-state hidden Markov models based on a Bayesian approach," in *Proc. ICSLP2002*, 2002, vol. 4, pp. 2669–2672.

[40] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Selection of shared-states hidden Markov model structure using Bayesian criterion," *IEICE Transactions on Information and Systems*, vol. J86-D-II, pp. 776–786, 2003, (in Japanese).

[41] T. Jitsuhiro and S. Nakamura, "Automatic generation of non-uniform HMM structures based on variational Bayesian approach," in *Proc. ICASSP2004*, 2004, vol. 1, pp. 805–808.

[42] K. Hashimoto, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "Bayesian context clustering using cross valid prior distribution for hmm-based speech recognition," in *Proc. Interspeech'08*, 2008.

[43] S. Shiota, K. Hashimoto, Y. Nankaku, and K. Tokuda, "Deterministic annealing based training algorithm for bayesian speech recognition," in *Proc. Interspeech' 09*, 2009, pp. 680–683.

[44] J.C. Chen and J.T. Chien, "Bayesian large margin hidden Markov models for speech recognition," in *Proc. ICASSP 2009*, 2009, pp. 3765–3768.

[45] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Application of variational Bayesian approach to speech recognition," in *Proc. Fall Meeting of ASJ 2002*, 2002, vol. 1, pp. 127–128, (in Japanese).

[46] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Bayesian acoustic modeling for spontaneous speech recognition," in *Proc. SSPR2003*, 2003, pp. 47–50.

[47] F. Valente and C. Wellekens, "Variational Bayesian GMM for speech recognition," in *Proc. Eurospeech2003*, 2003, pp. 441–444.

[48] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Application of variational Bayesian estimation and clustering to acoustic model adaptation," in *Proc. ICASSP2003*, 2003, vol. 1, pp. 568–571.

[49] S. Watanabe and A. Nakamura, "Acoustic model adaptation based on coarse-fine training of transfer vectors and its application to speaker adaptation task," in *Proc. ICSLP2004*, 2004, vol. 4, pp. 2933–2936.

[50] K. Yu and M. J. F. Gales, "Bayesian adaptation and adaptively trained systems," in *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU) 2005*, 2005, pp. 209–214.

[51] S. Watanabe and A. Nakamura, "Speech recognition based on Student's t-distribution derived from total Bayesian framework," *IEICE Transactions on Information and Systems*, vol. E89-D, pp. 970–980, 2006.

[52] S. Watanabe, A. Nakamura, and B.H. Juang, "Bayesian linear regression for hidden Markov model based on optimizing variational bounds," in *Proc. MLSP 2011*, 2011, (accepted).

[53] S. Watanabe, A. Sako, and A. Nakamura, "Automatic determination of acoustic model topology using variational Bayesian estimation and clustering," in *Proc. ICASSP2004*, 2004, vol. 1, pp. 813–816.

[54] F. Valente, "Infinite models for speaker clustering," in *Proc. Interspeech' 06*, 2006, pp. 1329–1332.

[55] N. Ding and Z. Ou, "Variational nonparametric bayesian hidden markov model," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 2098–2101.

[56] A. Ogawa and S. Takahashi, "Weighted distance measures for efficient reduction of gaussian mixture components in hmm-based acoustic model," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4173–4176.

[57] S. Watanabe and A. Nakamura, "Effects of Bayesian predictive classification using variational Bayesian posteriors for sparse training data in speech recognition," in *Proc. Interspeech '2005 - Eurospeech*, 2005, pp. 1105–1108.

[58] T. Mishina and M. Yamamoto, "Context adaptation using variational Bayesian learning for ngram models based on probabilistic LSA," *IEICE Transactions on Information and Systems*, vol. J87-D-II, pp. 1409–1417, 2004, (in Japanese).

[59] Y.-C. Tam and T. Schultz, "Dynamic language model adaptation using variational Bayes inference," in *Proc. Interspeech '2005 - Eurospeech*, 2005, pp. 5–8.

[60] J.T. Chien and C.H. Chueh, "Dirichlet class language models for speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 482–495, 2011.

[61] Y. Nankaku, H. Zen, K. Tokuda, T. Kitamura, and T. Masuko, "A bayesian approach to hmm-based speech synthesis," in *IEICE Technical Report*, 2003, vol. SP103(264), pp. 19–24, (in Japanese).

[62] K. Hashimoto, H. Zen, Y. Nankaku, T. Masuko, and K. Tokuda, "A bayesian approach to hmm-based speech synthesis," in *Proc, ICASSP 2009*, 2009, pp. 4029–4032.