SIP (2012), vol. 1, e5, page 1 of 11 © The Authors, 2012.

doi:10.1017/ATSIP.2012.6

OVERVIEW PAPER

Bayesian approaches to acoustic modeling: a review

SHINJI WATANABE[1] AND ATSUSHI NAKAMURA[2]

This paper focuses on applications of Bayesian approaches to acoustic modeling for speech recognition and related speech-processing applications. Bayesian approaches have been widely studied in the fields of statistics and machine learning, and one of their advantages is that their generalization capability is better than that of conventional approaches (e.g., maximum likelihood). On the other hand, since inference in Bayesian approaches involves integrals and expectations that are mathematically intractable in most cases and require heavy numerical computations, it is generally difficult to apply them to practical speech recognition problems. However, there have been many such attempts, and this paper aims to summarize these attempts to encourage further progress on Bayesian approaches in the speech-processing field. This paper describes various applications of Bayesian approaches to speech processing in terms of the four typical ways of approximating Bayesian inferences, i.e., maximum a posteriori *approximation, model complexity control using a Bayesian information criterion based on asymptotic approximation, variational approximation, and Markov chain Monte Carlo-based sampling techniques.*

Keywords: Speech processing, Machine learning, Bayesian approach, Approximate Bayesian inference

Received 8 February 2012; Revised 18 October 2012

I. INTRODUCTION

Speech recognition systems, which convert speech into text, make it possible for computers to process the information contained in human speech. The current successes in speech recognition and related speech-processing applications are based on pattern recognition that uses statistical learning theory. Maximum likelihood (ML) methods have become the standard techniques for constructing acoustic and language models for speech recognition. They guarantee that ML estimates approach the stationary values of the parameters. ML methods are also applicable to latent variable models, such as hidden Markov models (HMMs) and Gaussian mixture models (GMMs), thanks to the expectation–maximization (EM) algorithm [1]. Acoustic modeling based on HMMs and GMMs is one of the most successful examples of the ML–EM approach, and it has been greatly developed in previously reported studies [2–4].

However, the performance of current speech recognition systems is far from satisfactory. Specifically, the recognition performance is much poorer than the human capability of recognizing speech. This is because 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 in various factors such as the speaker, context, speaking style, and noise. For example, the performance of acoustic models trained using read speech degrades greatly when the models are used to recognize spontaneous speech due to the mismatch between the read and spontaneous speech characteristics [5]. More generally, 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. This paper addresses various attempts to improve the acoustic model training method beyond the conventional ML approach by employing *Bayesian* approaches.

In Bayesian approaches, all the variables that are introduced when models are parameterized, such as model parameters and latent variables, are regarded as probabilistic variables, and their posterior distributions are simply obtained by using the probabilistic sum and product rules. The difference between the Bayesian and ML approaches is that the estimation target is a probability distribution 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 [6–8]. However, the Bayesian approach requires complex integral and expectation computations

[1]Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge, MA 02139, USA.
[2]NTT Communication Science Laboratories, NTT Corporation, 2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan.

Corresponding author: Shinji Watanabe
Email: watanabe@merl.com

to obtain posterior distributions when models have latent variables. For example, to infer the posterior distribution of HMM/GMM model parameters $\mathbf{\Theta}$ given speech feature vectors $\mathbf{O}$, we need to calculate the following equation:

$$p(\mathbf{\Theta}|\mathbf{O}) = \sum_{\mathbf{Z}} \frac{p(\mathbf{O},\mathbf{Z}|\mathbf{\Theta})\,p(\mathbf{\Theta})}{p(\mathbf{O})}, \tag{1}$$

where $\mathbf{Z}$ is a set of HMM state and GMM component sequences. Once we obtain the posterior distribution, we classify category c (phoneme or word) given new speech feature vectors $\mathbf{x}$ based on the following posterior distribution:

$$p(c|\mathbf{x},\mathbf{O}) = \int p(c|\mathbf{\Theta},\mathbf{x})\,p(\mathbf{\Theta}|\mathbf{O})\,d\mathbf{\Theta}. \tag{2}$$

Since the integral and expectation often cannot be computed analytically, we need some approximations if we are to implement a Bayesian approach for a classification problem in speech processing.

There have already been many attempts to undertake Bayesian speech processing by approximating the above Bayesian inference [8, 9]. The most famous application of Bayesian approaches employs maximum *a posteriori* (MAP) approximation, which uses the maximum value of the posterior distribution instead of integrating out the latent variable or model parameter [7]. Historically, MAP-based speech recognition approaches constitute the first successful applications of Bayesian approaches to speech processing. These approaches were introduced in the early 1990s to deal with speaker adaptation problems in speech recognition [10, 11]. Around 1995, they started to be applied to more practical speech processing problems (e.g., continuous density HMM [12], which is a standard acoustic model in speech recognition, and speaker recognition based on a universal background model [13]). Other successful methods are based on the Bayesian information criterion (BIC), which is obtained by using asymptotic approximations [14, 15]. Starting around 2000, these methods have been applied to wide areas of speech processing, from phonetic decision tree clustering to speaker segmentation [16–19]. Recently, advanced Bayesian topics such as variational Bayes (VB) and Markov chain Monte Carlo (MCMC) have been actively studied in the machine learning field [8], and these approaches are also starting to be applied to speech processing [20–23], by following the successful Bayesian applications based on MAP and BIC.

Focusing on the four major trends as regards approximating Bayesian inferences, i.e., MAP approximation, asymptotic approximation for model complexity control, variational approximation, and MCMC, this paper aims to provide an overview of the various attempts described above in order to encourage researchers in the speech-processing field to investigate Bayesian approaches and guide them in this endeavor.

In addition to the above topics, there are other interesting Bayesian approaches that have been successfully applied to speech recognition, e.g., on-line Bayesian adaptation [24, 25], structural Bayes [26, 27], quasi-Bayes [28–30], graphical model representation [31–33], and Bayesian sensing HMM [34]. Although we do not focus on these approaches in detail, they have been summarized in other review and tutorial articles [35–37].

II. MAP

MAP approaches were introduced into speech recognition to utilize prior information [10–12]. The Bayesian approach is based on posterior distributions of the distribution parameters, while the ML approach only considers a particular value for these distribution parameters. Let $\mathbf{O} = \{\mathbf{o}_t \in \mathbb{R}^D | t = 1, \ldots, T\}$ be a given training dataset of D-dimensional feature vectors and $\mathbf{Z} = \{z_t | t = 1, \ldots, T\}$ be a set of corresponding latent variables. The posterior distribution for a distribution parameter Θ_c of category c is obtained by using the well-known Bayes theorem as follows:

$$p(\Theta_c|\mathbf{O},m) = \sum_{\mathbf{Z}} \int \frac{p(\mathbf{O},\mathbf{Z}|\Theta,m)\,p(\Theta|m)}{p(\mathbf{O}|m)} d\Theta_{-c}, \tag{3}$$

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 except c. In this paper, we regard the hyperparameter setting as the model structure, and include its variations in the index m. From equation (3), prior information can be utilized via estimations of the posterior distribution, which depends on prior distributions.

Equation (3) generally cannot be calculated analytically due to the summation over latent variables. To avoid the problem, MAP approaches approximate the distribution estimation as a point estimation. Namely, instead of obtaining the posterior distribution in equation (3), MAP approaches consider the following value

$$\begin{aligned} \Theta_c^{MAP} &= \operatorname*{argmax}_{\Theta_c} p(\Theta_c|\mathbf{O},m) \\ &= \operatorname*{argmax}_{\Theta_c} \sum_{\mathbf{Z}} p(\mathbf{O},\mathbf{Z}|\Theta_c,m)\,p(\Theta_c|m). \end{aligned} \tag{4}$$

This estimation can be efficiently performed by using the EM algorithm. The MAP approximation was first applied to the estimation of single-Gaussian HMM parameters in [10] and later extended to GMM–HMMs in [11, 12]. The effectiveness of MAP approaches can be illustrated in a speaker recognition task where prior distributions are set by speaker-independent HMMs. For example, Gauvain and Lee [12] compares speaker adaptation performance by employing ML and MAP estimations of acoustic model parameters using the DARPA Naval Resources Management (RM) task [38]. With 2 minutes of adaptation data, the ML word error rate was 31.5% and was worse than the speaker independent word error rate (13.9%) due to the over-training effect. However, the MAP word error

rate was 8.7%, clearly showing the effectiveness of the MAP approach. MAP estimation has also been used in speaker recognition based on universal background models [13], and in the discriminative training of acoustic models in speech recognition as a parameter-smoothing technique [39].

III. BIC

BIC approaches were introduced into speech recognition to perform model selection [16, 17]. To deal with model structure in a Bayesian approach, we can consider the following posterior distribution:

$$p(m|\mathbf{O}) = \sum_{\mathbf{Z}} \int \frac{p(\mathbf{O}, \mathbf{Z}|\Theta, m)\, p(\Theta|m)\, p(m)}{p(\mathbf{O})} d\Theta, \quad (5)$$

where $p(m)$ denotes a prior distribution for the model structure m. However, as with MAP approaches, equation (5) cannot be calculated analytically due to the summation over latent variables. The BIC only focuses on models that do not have latent variables. Under the asymptotic assumption (i.e., the assumption that there is a large amount of data), one can obtain the following equation:

$$\log p(m|\mathbf{O}) \propto \log p(\mathbf{O}|\Theta, m) - \frac{\#(\Theta)}{2} \log T. \quad (6)$$

The first term on 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, denoted by $\#(\Theta)$.

This criterion is widely used in speech processing. For example, it enables phonetic decision tree clustering to be performed in [16, 17] without having to set a heuristic stopping criterion as was done in [40]. Shinoda and Watanabe [16] show the effectiveness of the BIC/MDL[1] criterion for phonetic decision tree clustering in a 5000 Japanese word recognition task by comparing the performance of acoustic models based on BIC/MDL with models based on heuristic stopping criteria (namely, the state occupancy count and the likelihood threshold). BIC/MDL selected 2069 triphone HMM states automatically with an 80.4% recognition rate, while heuristic stopping criteria selected 1248 and 591 states with recognition rates of 77.9 and 66.6% in the best and worst cases, respectively. This result clearly shows the effectiveness of model selection using BIC/MDL. An extension of the BIC objective function by considering a tree structure is also discussed in [41], and an extension based on VB is discussed in Section IV. In addition, BIC/MDL is used for Gaussian pruning in acoustic models [19], and speaker segmentation [18]. BIC-based speaker segmentation is a particularly important technique for speaker diarization, which has been widely studied recently [42].

MAP and BIC, together with Bayesian Predictive Classification (BPC) [43, 44], which marginalizes model parameters so that the effect of over-training is mitigated and robust classification is obtained, can be practically realized in speech recognition. However, while Bayesian approaches can potentially have the three following advantages:

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

MAP, BIC, and BPC each have only one. In general, these advantages make pattern recognition methods more robust than those based on ML approaches. For example, a MAP-based framework approximates the posterior distribution of the parameter by using a MAP approximation to utilize prior information. BIC/MDL- and BPC-based frameworks, respectively, perform some sort of model selection and robust classification. These approaches are simple and powerful frameworks with which to transfer some of the advantages expected from Bayesian approaches to speech recognition systems. However, they also lose some of these advantages due to the approximations they introduce, as shown in Table 1. In the next section, we introduce another method for approximating a Bayesian inference, variational approximation, which includes all three Bayesian advantages simultaneously unlike the MAP, BIC, and BPC approaches.

Table 1. Comparison of VBEC and other Bayesian frameworks in terms of Bayesian advantages.

Bayesian advantage	VBEC	MAP	BIC/MDL	BPC
(1) Prior utilization	√	√	–	–
(2) Model selection	√	–	√	–
(3) Robust classification	√	–	–	√

IV. VB

This section presents an application of VB, a technique originally developed in the field of machine learning [45–48], to speech recognition. With this VB approach, approximate posterior distributions (VB posterior distributions) can be obtained effectively by iterative calculations similar to the EM algorithm used in the ML approach, while the three advantages of the Bayesian approaches are retained. Therefore, the framework is formulated using VB to replace the ML approaches with Bayesian approaches in speech recognition. We briefly review a speech recognition framework based on a fully Bayesian approach to overcome the lack of robustness described above by utilizing the three Bayesian advantages [20, 21]. A detailed discussion of the formulation and experiments can be found in [49].

A) Application of VB to speech recognition

As we saw earlier, Bayesian approaches aim at obtaining posterior distributions for the model parameters, but these

[1]BIC and minimum description length (MDL) criteria have been independently proposed, but they are practically the same. Therefore, they are identified in this paper and referred to as BIC/MDL.

posterior distributions cannot be generally obtained analytically. The goal of VB is to approximate these posterior distributions using some other distributions, referred to as variational distributions, which are optimized so that they are as close as possible, in some sense yet to be defined, to the true posterior distributions. The variational distributions are generally assumed to belong to a family of distributions of a simpler form than the original posterior distributions. Here, we consider an arbitrary posterior distribution q, and assume that it can be factorized as

$$q(\Theta, \mathbf{Z}, m|\mathbf{O}) = \prod_c q(\Theta_c|\mathbf{O}_c, m) q(\mathbf{Z}_c|\mathbf{O}_c, m) q(m|\mathbf{O}_c), \tag{7}$$

where c is a category index (e.g., a phoneme if we deal with a phoneme-based acoustic model). VB then focuses on minimizing the Kullback–Leibler divergence from $q(\Theta, \mathbf{Z}, m|\mathbf{O})$ to $p(\Theta, \mathbf{Z}, m|\mathbf{O})$, which can be shown to be equivalent to maximizing the following objective functional:

$$\mathcal{F}^m[q(\Theta_c|\mathbf{O}_c, m), q(\mathbf{Z}_c|\mathbf{O}_c, m)] = \left\langle \log \frac{p(\mathbf{O}_c, \mathbf{Z}_c|\Theta_c, m)\, p(\Theta_c|m)}{q(\Theta_c|\mathbf{O}_c, m) q(\mathbf{Z}_c|\mathbf{O}_c, m)} \right\rangle_{q(\Theta_c|\mathbf{O}_c, m), q(\mathbf{Z}_c|\mathbf{O}_c, m)}, \tag{8}$$

where 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. Equation (8) can be shown to be a lower bound of the marginalized log likelihood. The optimal posterior distribution can be obtained by a variational method, which due to the factorization assumption (7) leads to:

$$\begin{aligned}
\widetilde{q}(\Theta_c|\mathbf{O}_c, m) &= \operatorname*{argmax}_{q(\Theta_c|\mathbf{O}, m)} \mathcal{F}^m[q(\Theta_c|\mathbf{O}_c, m), q(\mathbf{Z}_c|\mathbf{O}_c, m)], \\
\widetilde{q}(\mathbf{Z}_c|\mathbf{O}_c, m) &= \operatorname*{argmax}_{q(\mathbf{Z}_c|\mathbf{O}, m)} \mathcal{F}^m[q(\Theta_c|\mathbf{O}_c, m), q(\mathbf{Z}_c|\mathbf{O}_c, m)], \\
\widetilde{q}(m|\mathbf{O}) &= \operatorname*{argmax}_{q(m|\mathbf{O})} \sum_c \mathcal{F}^m[q(\Theta_c|\mathbf{O}_c, m), q(\mathbf{Z}_c|\mathbf{O}_c, m)].
\end{aligned} \tag{9}$$

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 where the MAP probability can be selected as follows:

$$\widetilde{m} = \operatorname*{argmax}_{\{m\}} \widetilde{q}(m|\mathbf{O}) = \operatorname*{argmax}_{\{m\}} \mathcal{F}^m. \tag{10}$$

This indicates that by maximizing the total $\mathcal{F}^m$ with respect to not only $q(\Theta_c|\mathbf{O}_c, m)$ and $q(Z_c|\mathbf{O}_c, m)$ but also m, we can obtain the optimal parameter distributions and can select the optimal model structure simultaneously [47, 48]. The VB approach is applied to a continuous density HMM (left-to-right HMM with a GMM for each state) in the variational Bayesian estimation and clustering (VBEC) for speech recognition framework [20, 21]. The continuous density HMM is a standard acoustic model that represents a phoneme category for speech recognition. VBEC is a fully Bayesian framework, where all the following acoustic model procedures for speech recognition (acoustic model construction and speech classification) are re-formulated in a VB manner:

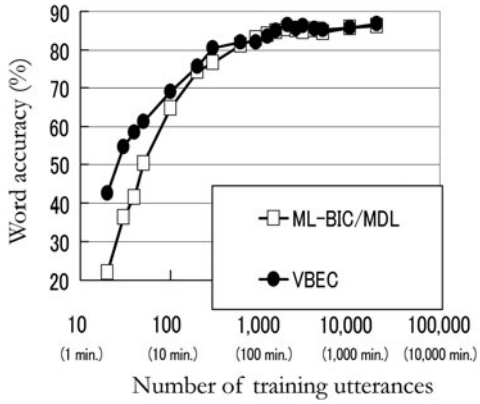

Fig. 1. Superiority of VBEC-based acoustic model construction for a small amount of training data.

- Output distribution setting
  → Output and prior distribution setting
- Parameter estimation by ML Baum–Welch
  → Posterior estimation by VB Baum–Welch
- Model selection by using heuristics
  → Model selection by using variational lower bound
- Classification using ML estimates
  → BPC using VB posteriors

Consequently, VBEC includes the three Bayesian advantages unlike the conventional Bayesian approaches, as illustrated in Table 1.

B) Experiments and related work

We briefly illustrate the effectiveness of the VBEC framework using the results of speech recognition experiments (see [49] for details). Figure 1 compares word accuracies on Japanese read speech data (JNAS) for various amounts of training data used in acoustic model construction. The difference between VBEC and conventional ML- and BIC/MDL-based acoustic modeling is whether or not the approach utilizes prior distributions. VBEC significantly improved the performance for a small amount of training data, which shows the effectiveness of (1) a prior utilization function in Bayesian approaches. Table 2 shows experimental results for the automatic determination of the acoustic model topology by using VBEC and the conventional heuristic approach that determines the model topology by evaluating ASR performance on development sets. In the various ASR tasks, VBEC obtained comparable performance to the conventional method by selecting appropriate model topologies without using a development set, which shows the effectiveness of (2) a model selection function in Bayesian approaches. Finally, Fig. 2 shows a comparison of word accuracies with Corpus of Spontaneous Japanese (CSJ) data [5] in speaker adaptation experiments.

Table 2. Automatic determination of acoustic model topology.

	Japanese read speech (JNAS)	Japanese isolated word (JEIDA)	Japanese lecture (CSJ)	English read speech (WSJ)
VBEC (# states, # components)	91.7 % (912, 40)	97.9 % (254, 35)	74.5 % (1986, 32)	91.3 % (2504, 32)
ML + dev. Set (# states, # components)	91.4 % (1000, 30)	98.1 % (1000, 15)	74.2 % (3000, 32)	91.3 % (7500, 32)

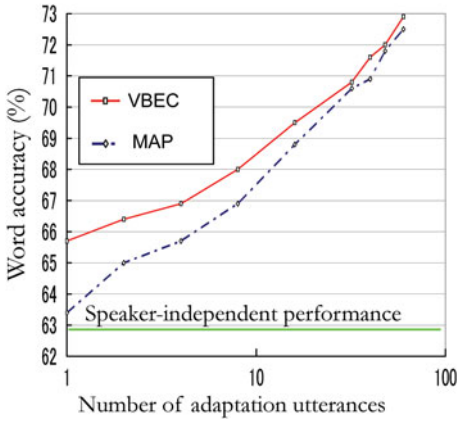

Fig. 2. Robust classification based on marginalization effect.

Table 3. Technical trend of speech recognition using VB

Topic	References
Feature extraction	[50, 51]
Speech GMM for noise robust ASR and voice activity detection	[52, 53]
Formulation of Bayesian speech recognition	[20, 21, 54, 55]
Selection of number of GMM components	[56–58]
Acoustic model adaptation	[59–62]
Determination of acoustic model topology	[63–67]
Non-parametric Bayes for acoustic models/speaker diarization	[68–71]
Statistical speech synthesis	[72]

VBEC and MAP used the same prior distributions, and the difference between them is whether or not the model parameters are marginalized (integrated out). VBEC also significantly improved the performance for a small amount of training data, which shows the effectiveness of (3) a robust classification function in Bayesian approaches. Thus, these results confirm experimentally that VBEC includes the three Bayesian advantages unlike the conventional Bayesian approaches, as shown in Table 1.

VB is becoming a common technique in speech processing. Table 3 summarizes the technical trend in speech processing techniques involving VB. Note that VB has been widely applied to speech recognition and other forms of speech processing. Given such a trend, VBEC is playing 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 to the approximation of Bayesian inferences, the variational techniques are used as an effective approximation method in some speech processing problems, e.g., approximating the Kullback–Leibler divergence between GMMs [73], and the Bayesian treatment of a discriminative HMM by using minimum relative entropy discrimination [74].

V. MCMC

In previous sections, we described Bayesian approaches based on deterministic approximations (MAP, asymptotic approximation, and VB). Another powerful way to implement Bayesian approaches is to rely on a sampling method, which obtains expectations by using Monte Carlo techniques [7, 8]. The main advantage of the sampling approaches is that they can avoid local optimum problems in addition to providing other Bayesian advantages (mitigation of data sparseness problems and capacity for model structure optimization). While their heavy computational cost could be a problem in practice, recent improvements in computational power and the development of theoretical and practical aspects have allowed researchers to start applying them to practical problems (e.g., [75, 76] in natural language processing). This paper describes our recent attempts to apply a sampling approach to acoustic modeling based on MCMC, in particular Gibbs sampling [23, 71, 77]. Gibbs sampling is a simple and widely applicable sampling algorithm [78] that samples the latent variable z_t by using the conditional distribution $p(z_t|\mathbf{z}_{\backslash t})$ where $\mathbf{z}_{\backslash t}$ is the set of all latent variables except z_t. By iteratively sampling z_t for all t based on this conditional distribution, we can efficiently sample the latent variables, which are then used to compute the expectations (e.g., equation (1)) required in Bayesian approaches. Here, we focus on an example of a hierarchical GMM, called a multi-scale mixture model, used as an acoustic model in speaker clustering, and introduce a formulation based on Gibbs sampling.

A) Formulation

Multi-scale mixture model (M^3)

M^3 considers two types of observation vector sequences. One is an utterance- (or segment-) level sequence and the

other is a frame-level sequence. A D-dimensional observation vector (e.g., MFCC) at frame t in utterance u is represented as $\mathbf{o}_{u,t} (\in \mathbb{R}^D)$. A set of observation vectors in utterance u is represented as $\mathbf{o}_u \triangleq \{\mathbf{o}_{u,t}\}_{t=1}^{T_u}$.

We assume that the frame-level sequence is modeled by a GMM as usual, and the utterance-level sequence is modeled by a mixture of these GMMs. Two kinds of latent variables are involved in M^3 for each sequence: utterance-level latent variables z_u and frame-level latent variables $v_{u,t}$. Utterance-level latent variables may represent emotion, topic, and speaking style as well as speakers, depending on the speech variation. The likelihood function of U observation vectors ($\mathbf{O} \triangleq \{\mathbf{o}_u\}_{u=1}^{U}$) given the latent variable sequences ($\mathbf{Z} \triangleq \{z_u\}_u$ and $\mathbf{V} \triangleq \{v_{u,t}\}_{u,t}$) can be expressed as follows:

$$p(\mathbf{O}|\mathbf{Z},\mathbf{V},\boldsymbol{\Theta}) = \prod_{u=1}^{U} h_{z_u} \prod_{t=1}^{T_u} w_{z_u,v_{u,t}} \mathcal{N}(\mathbf{o}_{u,t}|\boldsymbol{\mu}_{z_u,v_{u,t}}, \boldsymbol{\Sigma}_{z_u,v_{u,t}}), \tag{11}$$

where $\{h_s\}_s, \{w_{s,k}\}_{s,k}, \{\boldsymbol{\mu}_{s,k}\}_{s,k}, \{\boldsymbol{\Sigma}_{s,k}\}_{s,k} (\triangleq \boldsymbol{\Theta})$ are the utterance-level mixture weight, frame-level mixture weight, mean vector, and covariance matrix parameters, respectively. s and k denote utterance-level and frame-level mixture indexes, respectively. $\mathcal{N}$ denotes a normal distribution.

Let us now consider the Bayesian treatment of this multi-scale mixture model. We assume a diagonal covariance matrix for the Gaussian distributions as usual, where the d–d diagonal element of the covariance matrix is expressed as σ_{dd}, and use the following conjugate distributions as the prior distributions of the model parameters:

$$p(\boldsymbol{\Theta}|\boldsymbol{\Psi}^0) = \left\{ \begin{array}{ll} \mathbf{h} & \sim \mathcal{D}(\mathbf{h}^0) \\ \mathbf{w}_s & \sim \mathcal{D}(\mathbf{w}^0) \\ \boldsymbol{\mu}_{s,k} & \sim \mathcal{N}(\boldsymbol{\mu}_k^0, (\xi^0)^{-1}\boldsymbol{\Sigma}_{s,k}) \\ (\sigma_{s,k,dd})^{-1} & \sim \mathcal{G}(\eta^0, \sigma_{k,dd}^0) \end{array} \right\}, \tag{12}$$

where $\mathbf{h}^0, \mathbf{w}^0, \boldsymbol{\mu}_k^0, \xi^0, \sigma_{k,dd}^0, \eta^0 (\triangleq \boldsymbol{\Psi}^0)$ are the hyperparameters. $\mathcal{D}$ and $\mathcal{G}$ denote Dirichlet and Gamma distributions, respectively. The generative process of M^3 is shown in Fig. 3. Based on the generative model, we derive analytical solutions for Gibbs samplers of the multi-scale mixture model based on the marginalized likelihood for the complete data.

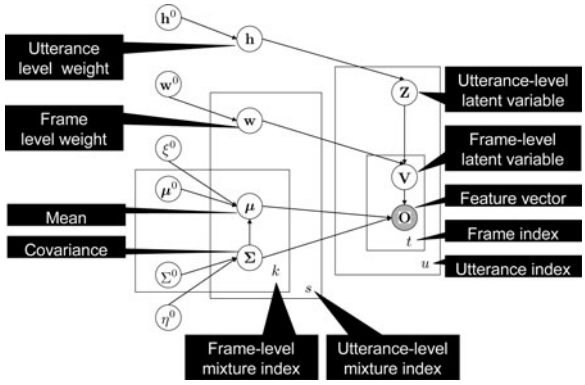

Fig. 3. Graphical representation of multi-scale mixture model.

Gibbs sampler

Frame-level mixture component

The function form of the Gibbs sampler, which assigns frame-level mixture component k at frame t probabilistically, is analytically obtained as follows:

$$p(v_{u,t} = k'|\mathbf{O},\mathbf{V}_{\backslash t},\mathbf{Z}_{\backslash u}, z_u = s) = \frac{\exp\left(g_{s,k'}(\tilde{\boldsymbol{\Psi}}_{s,k'}) - g_{s,k'}(\tilde{\boldsymbol{\Psi}}_{s,k'\backslash t})\right)}{\sum_k \exp\left(g_{s,k}(\tilde{\boldsymbol{\Psi}}_{s,k}) - g_{s,k}(\tilde{\boldsymbol{\Psi}}_{s,k\backslash t})\right)}. \tag{13}$$

Here, $\mathbf{O}_{\backslash t}$ and $\mathbf{V}_{\backslash t}$ indicate sets that do not include the tth frame elements. $\mathbf{Z}_{\backslash u}$ indicates a set that does not include the uth utterance element. $\tilde{\boldsymbol{\Psi}}_{s,k\backslash t}$ is computed by the sufficient statistics using $\mathbf{O}_{\backslash t}$ and $\mathbf{V}_{\backslash t}$. $g_{s,k}(\cdot)$ is defined as follows:

$$g_{s,k}(\tilde{\boldsymbol{\Psi}}_{s,k}) \triangleq \log\Gamma(\tilde{w}_{s,k}) - \frac{D}{2}\log\tilde{\xi}_{s,k} + D\log\Gamma\left(\frac{\tilde{\eta}_{s,k}}{2}\right) - \frac{\tilde{\eta}_{s,k}}{2}\sum_d \log\tilde{\sigma}_{s,k,dd},$$

where $\tilde{\mathbf{h}}_s, \tilde{\mathbf{w}}_s, \tilde{\boldsymbol{\mu}}_{s,k}, \tilde{\xi}_{s,k}, \tilde{\sigma}_{s,k,dd}$ and $\tilde{\eta}_{s,k} (\triangleq \tilde{\boldsymbol{\Psi}})$ are the hyperparameters of the posterior distributions for $\boldsymbol{\Theta}$, which are obtained from the hyperparameters of the prior distributions ($\boldsymbol{\Psi}^0$) and the sufficient statistics as follows:

$$\left\{ \begin{array}{ll} \tilde{h}_s & = h_s^0 + c_s, \\ \tilde{w}_{s,k} & = w_k^0 + n_{s,k}, \\ \tilde{\xi}_{s,k} & = \xi^0 + n_{s,k}, \\ \tilde{\mu}_{s,k} & = \dfrac{\xi^0\boldsymbol{\mu}_k^0 + \mathbf{m}_{s,k}}{\tilde{\xi}_{s,k}}, \\ \tilde{\eta}_{s,k} & = \eta^0 + n_{s,k}, \\ \tilde{\sigma}_{s,k,dd} & = \sigma_{k,dd}^0 + r_{s,k,dd} + \xi^0(\mu_{k,d}^0)^2 - \tilde{\xi}_{s,k}(\tilde{\mu}_{s,k,d})^2. \end{array} \right. \tag{14}$$

c_s is the count of utterances assigned to s and $n_{s,k}$ is the count of frames assigned to k in s. $\mathbf{m}_{s,k}$ and $r_{s,k,dd}$ are first-order and second-order sufficient statistics, respectively.

Utterance-level mixture component

As with the frame-level mixture component case, the Gibbs sampler assigns utterance-level mixture s at utterance u by using the following equation:

$$\log p(z_u = s|\mathbf{O},\mathbf{V},\mathbf{Z}_{\backslash u}) \propto \log\frac{\Gamma(\sum_k \tilde{w}_{s\backslash u,k})}{\Gamma(\sum_k \tilde{w}_{s,k})} + \sum_k g_{s,k}(\tilde{\boldsymbol{\Psi}}_{s,k}) - g_{s,k}(\tilde{\boldsymbol{\Psi}}_{s\backslash u,k}).$$

$\mathbf{O}_{\backslash u}$ and $\mathbf{V}_{\backslash u}$ indicate sets that do not include subsets of the frame elements in u. $\tilde{\boldsymbol{\Psi}}_{s\backslash u,k}$ is computed by the sufficient statistics using $\mathbf{O}_{\backslash u}$ and $\mathbf{V}_{\backslash u}$. Therefore, the posterior

Algorithm 1 Gibbs sampling based multi-scale mixture model.

1: Initialize $\mathbf{\Phi}^0$
2: **repeat**
3: **for** $u =$ shuffle $(1 \cdots U)$ **do**
4: **for** $t =$ shuffle $(1 \cdots T_u)$ **do**
5: Sample $v_{u,t}$ by using equation (13)
6: **end for**
7: **end for**
8: **for** $u =$ shuffle $(1 \cdots U)$ **do**
9: Sample z_u by using equation (15)
10: **end for**
11: **until** some condition is met

probability can be obtained as follows:

$$p(z_u = s' | \mathbf{O}, \mathbf{V}, \mathbf{Z}_{\backslash u}) = \frac{\exp\left(\log \frac{\Gamma(\sum_k \tilde{w}_{s'\backslash u,k})}{\Gamma(\sum_k \tilde{w}_{s',k})} + \sum_k g_{s',k}(\tilde{\mathbf{\Psi}}_{s',k}) - g_{s',k}(\tilde{\mathbf{\Psi}}_{s'\backslash u,k})\right)}{\sum_{s,k} \exp\left(\log \frac{\Gamma(\sum_k \tilde{w}_{s\backslash u,k})}{\Gamma(\sum_k \tilde{w}_{s,k})} + g_{s,k}(\tilde{\mathbf{\Psi}}_{s,k}) - g_{s,k}(\tilde{\mathbf{\Psi}}_{s\backslash u,k})\right)}. \quad (15)$$

These solutions for the multi-scale mixture model based on Gibbs sampling jointly infer the latent variables by interleaving frame-level and utterance-level samples.

Algorithm 1 provides a sample code of the multi-scale mixture model.

B) Experiments

We describe experimental results obtained with the multi-scale mixture model for meeting data, recorded by NTT Communication Science Laboratories to analyze and recognize meetings [79]. We used four of the sessions (3402 utterances) to construct a prior GMM in advance, and the other two sessions as development (495 utterances spoken by four speakers), and evaluation sets (560 utterances spoken by four speakers), respectively. As an observation vector, we used MFCC features with log energy, Δ, and $\Delta\Delta$ components. As a preliminary experiment, the numbers of clusters were set at the correct answer. First, a prior GMM (i.e., a universal background model) was estimated by using the four sessions consisting of 3402 utterances based on the conventional ML–EM algorithm, and the values of the GMM parameters were set as those of the hyperparameters in M^3 $(\mathbf{w}^0, \boldsymbol{\mu}_k^0, \mathbf{\Sigma}_k^0)$. Figure 4 shows the speaker clustering performance of the multi-scale mixture (M^3 Gibbs), the MAP-based approach (M^3 MAP–EM) and the conventional BIC-based approach in terms of the frame-level error rate of each method based on the diarization error rate defined by NIST [80]. Speaker clustering experiments showed that M^3 Gibbs provided a significant improvement over the conventional BIC and M^3 MAP–EM-based approaches. The main advantage of M^3 Gibbs and M^3 MAP–EM over BIC is that they can precisely model speaker clusters based on the GMM unlike the single Gaussian model used in BIC. In addition, M^3 Gibbs further improved on the speaker clustering performance of M^3 MAP–EM because the Gibbs sampling algorithm can avoid local optimum solutions unlike the MAP–EM algorithm. These superior characteristics are derived from the Gibbs-based Bayesian properties.

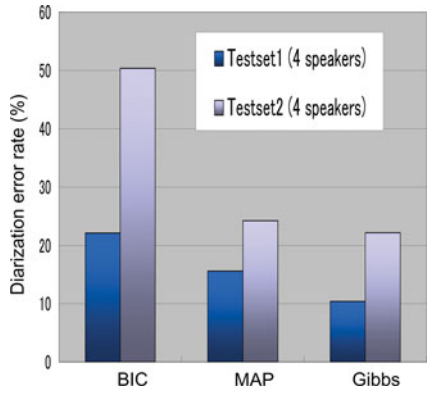

Fig. 4. Diarization error rate for NTT meeting data.

Table 4. Comparison of MCMC and VB for speaker clustering

Evaluation data	Method	ACP	ASP	K value
CSJ-1	MCMC	0.808	0.898	0.851
(# spkr10, # utt 50)	VB	0.704	0.860	0.777
CSJ-2	MCMC	0.852	0.892	0.871
(# spkr10, # utt 100)	VB	0.695	0.846	0.782
CSJ-3	MCMC	0.866	0.892	0.879
(# spkr10, # utt 200)	VB	0.780	0.870	0.823
CSJ-4	MCMC	0.784	0.694	0.738
(# spkr10, # utt 2,491)	VB	0.773	0.673	0.721
CSJ-5	MCMC	0.740	0.627	0.681
(# spkr10, # utt 2,321)	VB	0.693	0.676	0.684

MCMC-based acoustic modeling for speaker clustering was further investigated with respect to the difference in the MCMC and VB estimation methods by [71]. Table 4 shows speaker clustering results in terms of the average cluster purity (ACP), average speaker purity (ASP), and geometric mean of those values (K value) to the evaluation criteria in the speaker clustering. We used the Corpus of Spontaneous Japanese (CSJ) dataset [5] and investigated the speaker clustering performance for MCMC and VB for various amounts of data. Table 4 showed that the MCMC-based method outperformed the VB method by avoiding local optimum solutions, especially when only few utterances could be used. These results also supported the importance derived from the Gibbs-based Bayesian properties.

VI. SUMMARY AND FUTURE PERSPECTIVE

This paper introduced selected topics regarding Bayesian applications to acoustic modeling in speech processing.

As standard techniques, we first explained MAP- and BIC-based approaches. We then focused on applications of VB and MCMC, following the recent trend of Bayesian applications to speech recognition emphasizing the advantages of fully Bayesian approaches that explicitly obtain posterior distributions of model parameters and structures based on these two methods. These approaches are associated with the progress of Bayesian approaches in the statistics and machine learning fields, and speech recognition based on Bayesian approaches is likely to advance further, thanks to the recent progress in these fields.

One promising example of further progress is structure learning by using Bayesian approaches. This paper introduced a powerful advantage of Bayesian model selection for the structure learning of standard acoustic models in Sections III and IV. Furthermore, the recent success of deep learning for acoustic modeling [81] places more importance on the structure learning of deep network topologies (e.g., number of layers and number of hidden states) in addition to the conventional HMM topologies. To deal with the problem, advanced structure learning techniques based on non-parametric Bayes [82] would be a powerful candidate. These approaches have recently been actively studied in the machine-learning field [83–85]. In conjunction with this trend, various applications of non-parametric Bayes have been proposed in speech processing [22, 23, 86], spoken language processing [75, 76, 87], and music signal processing [88–90].

Another important future work is how to involve Bayesian approaches with discriminative approaches theoretically and practically, since discriminative training [39, 91], structured discriminative models [92], and deep discriminative learning [81] have become standard approaches in acoustic modeling. One promising approach for this direction is the marginalization of model parameters and margin variables to provide Bayesian interpretations with discriminative methods [93]. However, applying [93] to acoustic models requires some extensions to deal with large-scale structured data problems [74]. This extension enables the more robust regularization of discriminative approaches, and allows structure learning by combining Bayesian and discriminative criteria.

Finally, we believe that further progress based on Bayesian approaches for acoustic models would improve the success of speech processing applications including speech recognition. To this end, we encourage people in a wide range of research areas (e.g., speech processing, machine learning, and statistics) to explore this exciting and interdisciplinary topic.

ACKNOWLEDGMENTS

The authors thank Dr Jonathan Le Roux at Mitsubishi Electric Research Laboratories (MERL) for fruitful discussions. We also thank the anonymous reviewers for their valuable comments on our paper, which have improved its quality.

REFERENCES

[1] Dempster, A.P.; Laird, N.M.; Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B, **39**, (1976) 1–38.

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

[3] Huang, X.D.; Ariki, Y.; Jack, M.A.: Hidden Markov Models for Speech Recognition, *Edinburgh University Press*, 1990.

[4] Gales, M.; Young, S.: The application of hidden Markov models in speech recognition. Signal Process., **1**, (3), (2007) 195–304.

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

[6] Berger, J.O.: Statistical Decision Theory and Bayesian Analysis, 2nd ed., *Springer-Verlag*, 1985.

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

[8] Bishop, C.M.: Pattern Recognition and Machine Learning, vol. 4, *Springer* New York, 2006.

[9] Ghahramani, Z.: Unsupervised learning. Advanced Lectures on Machine Learning, 2004, 72–112, Springer.

[10] Lee, C.-H.; Lin, C.H.; Juang, B-H.: A study on speaker adaptation of the parameters of continuous density hidden Markov models. IEEE Trans. Acoust. Speech Signal Process., **39**, (1991) 806–814.

[11] Gauvain, J.L.; Lee, C.H.: Improved acoustic modeling with Bayesian learning. in *ICASSP'92*, **1**, (1992) 481–484.

[12] Gauvain, J.-L.; Lee, C.-H.: Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech Audio Process., **2**, (1994) 291–298.

[13] Reynolds, D.A.; Quatieri, T.F.; Dunn, R.B.: Speaker verification using adapted Gaussian mixture models. Digit. Signal Process., **10**, (1–3), (2000) 19–41.

[14] Schwarz, G.: Estimating the dimension of a model. Ann. Stat., **6**, (1978) 461–464.

[15] Akaike, H.: Likelihood and the Bayes procedure. in *Bayesian Statistics*, J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds. 1980, 143–166, *University Press*, Valencia, Spain.

[16] Shinoda, K.; Watanabe, T.: MDL-based context-dependent subword modeling for speech recognition. J. Acoust. Soc. Jpn. (E), **21**, (2000) 79–86.

[17] Chou, W.; Reichl, W.: Decision tree state tying based on penalized Bayesian information criterion, in *Proc. ICASSP1999*, **1**, (1999) 345–348.

[18] Chen, S.; Gopinath, R.: Model selection in acoustic modeling. in *Proc. Eurospeech1999*, **3**, (1999) 1087–1090.

[19] Shinoda, K.; Iso, K.: Efficient reduction of Gaussian components using MDL criterion for HMM-based speech recognition. in *Proc. ICASSP2001*, **1**, (2001) 869–872.

[20] Watanabe, S.; Minami, Y.; Nakamura, A.; Ueda, N.: Application of Variational Bayesian Approach to Speech Recognition, NIPS 2002, *MIT Press*, 2002, 1261–1268.

[21] Watanabe, S.; Minami, Y.; Nakamura, A.; Ueda, N.: Variational Bayesian estimation and clustering for speech recognition. IEEE Trans. Speech Audio Process., **12**, (2004) 365–381.

[22] Fox, E.B.; Sudderth, E.B.; Jordan, M.I.; Willsky, A.S.: An HDP-HMM for systems with state persistence. in *Proc. of ICML*, 2008, 312–319.

[23] Tawara, N.; Watanabe, S.; Ogawa, T.; Kobayashi, T.: Speaker clustering based on utterance-oriented Dirichlet process mixture model. in *Proc. Interspeech'11*, 2011, 2905–2908.

[24] Huo, Q.; Lee, C.-H.: On-line adaptive learning of the correlated continuous density hidden Markov models for speech recognition. IEEE Trans. Speech Audio Process., **6**, (1998) 386–397.

[25] Watanabe, S.; Nakamura, A.: Predictor–corrector adaptation by using time evolution system with macroscopic time scale. IEEE Trans. Audio Speech Lang. Process., **18**, (2), (2010) 395–406.

[26] Shinoda, K.; Lee, C.-H.: A structural Bayes approach to speaker adaptation. IEEE Trans. Speech Audio Process., **9**, (2001) 276–287.

[27] Siohan, O.; Myrvoll, T.A.; Lee, C.H.: Structural maximum a posteriori linear regression for fast HMM adaptation. Comput. Speech Lang., **16**, (1), (2002) 5–24.

[28] Makov, U.E.; Smith, A.F.M.: A quasi-Bayes unsupervised learning procedure for priors. IEEE Trans. Inf. Theory, **23**, (1977) 761–764.

[29] Huo, Q.; Chan, C.; Lee, C.-H.: On-line adaptation of the SCHMM parameters based on the segmental quasi-Bayes learning for speech recognition. IEEE Trans. Speech Audio Process., **4**, (1996) 141–144.

[30] Chien, J.T.: Quasi-Bayes linear regression for sequential learning of hidden Markov models. IEEE Trans. Speech Audio Process., **10**, (2002) 268–278.

[31] Zweig, G.; Russell, S.: Speech recognition with dynamic Bayesian networks, in *Proc. Nat. Conf. Artificial Intelligence*, 1998, 173–180.

[32] Bilmes, J.; Zweig, G.: The Graphical Models Toolkit: An open source software system for speech and time-series processing, in *Proc. ICASSP'02*, 2002, vol. 4, 3916–3919.

[33] Rennie, S.; Hershey, J.R.; Olsen, P.A.: Single channel multi-talker speech recognition: Graphical modeling approaches. IEEE Signal Process. Mag. Spec. Issue Graph. Models, **27**, (6), (2010) 66–80.

[34] Saon, G.; Chien, J.T.: Bayesian sensing hidden Markov models for speech recognition. in *Proc. ICASSP'11*. IEEE, 2011, 5056–5059.

[35] Lee, C.-H.; Huo, Q.: On adaptive decision rules and decision parameter adaptation for automatic speech recognition. in *Proc. IEEE*, **88**, (2000) 1241–1269.

[36] Bilmes, J.; Bartels, C.: Graphical model architectures for speech recognition. IEEE Signal Process. Mag., **22**, (5), (2005) 89–100.

[37] Watanabe, S.; Chien, J.T.: Tutorial: Bayesian learning for speech and language processing. T-10, ICASSP'12, 2012.

[38] Price, P.; Fisher, W.M.; Bernstein, J.; Pallett, D.S.: The DARPA 1000-word resource management database for continuous speech recognition, in *Proc. ICASSP'88*, 1988, 651–654.

[39] Povey, D.: *Discriminative training for large vocabulary speech recognition*, Ph.D. thesis, Cambridge University, 2003.

[40] Young, S.J.; Odell, JJ; Woodland, PC: Tree-based state tying for high accuracy acoustic modelling, in *Proc. Workshop on Human Language Technology*, 1994, 307–312.

[41] Hu, R.; Zhao, Y.: Knowledge-based adaptive decision tree state tying for conversational speech recognition. IEEE Trans. Audio Speech Lang. Process., **15**, (7), (2007) 2160–2168.

[42] Anguera Miro, X.; Bozonnet, S.; Evans, N.; Fredouille, C.; Friedland, G.; Vinyals, O.: Speaker diarization: A review of recent research. IEEE Trans. Audio Speech Lang. Process., **20**, (2), (2012) 356–370.

[43] Jiang, H.; Hirose, K.; Huo, Q.: Robust speech recognition based on a Bayesian prediction approach. IEEE Trans. Speech Audio Process., **7**, (1999) 426–440.

[44] Huo, Q.; Lee, C.-H.: A Bayesian predictive classification approach to robust speech recognition. IEEE Trans. Speech Audio Process., **8**, (2000) 200–204.

[45] Jordan, M.I.; Ghahramani, Z.; Jaakkola, T.S.; Saul, L.K.: An introduction to variational methods for graphical models. *Mach. Learn.*, **37**, (1997) 183–233.

[46] Waterhouse, S.; MacKay, D.; Robinson, T.: Bayesian Methods for Mixtures of Experts, NIPS 7, *MIT Press*, 1995, 351–357.

[47] Attias, H.: Inferring parameters and structure of latent variable models by variational Bayes, in *Proc. Uncertainty in Artificial Intelligence (UAI) 15*, 1999, 21–30.

[48] Ueda, N.; Ghahramani, Z.: Bayesian model search for mixture models based on optimizing variational bounds. Neural Netw., **15**, (2002) 1223–1241.

[49] Watanabe, S.: *Speech recognition based on a Bayesian approach*, Ph.D. thesis, Waseda University, 2006.

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

[51] Valente, F.; Wellekens, C.: Variational Bayesian feature selection for Gaussian mixture models, in *Proc. ICASSP2004*, 2004, vol. 1, 513–516.

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

[53] Cournapeau, D.; Watanabe, S.; Nakamura, A.; Kawahara, T.: Online unsupervised classification with model comparison in the variational Bayes framework for voice activity detection. *IEEE J. Sel. Top. Signal Process.*, **4**, (6), (2010) 1071–1083.

[54] Zhang, Y.; Liu, P.; Chien, J.T.; Soong, F.: An evidence framework for Bayesian learning of continuous-density hidden Markov models, in *Proc. ICASSP 2009*, 2009, 3857–3860.

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

[56] Watanabe, S.; Minami, Y.; Nakamura, A.; Ueda, N.: Bayesian acoustic modeling for spontaneous speech recognition, in *Proc. SSPR2003*, 2003, 47–50.

[57] Valente, F.; Wellekens, C.: Variational Bayesian GMM for speech recognition, in *Proc. Eurospeech2003*, 2003, 441–444.

[58] Ogawa, A.; Takahashi, S.: Weighted distance measures for efficient reduction of Gaussian mixture components in HMM-based acoustic model, in *Proc. ICASSP'08*, 2008, 4173–4176.

[59] Watanabe, S.; Nakamura, A.: 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.

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

[61] Watanabe, S.; Nakamura, A.; Juang, B.H.: Bayesian linear regression for hidden Markov model based on optimizing variational bounds,' in *Proc. MLSP 2011*, 2011, 1–6.

[62] Hahm, S.J.; Ogawa, A.; Fujimoto, M.; Hori, T.; Nakamura, A.: Speaker adaptation using variational Bayesian linear regression in normalized feature space, in *Proc. of Interspeech'12*, 2012.

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

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

[65] Watanabe, S.; Sako, A.; Nakamura, A.: Automatic determination of acoustic model topology using variational Bayesian estimation and

clustering for large vocabulary continuous speech recognition. IEEE Trans. Audio Speech Lang. Process. **14**, (2006) 855–872.

[66] Hashimoto, K.; Zen, H.; Nankaku, Y.; Lee, A.; Tokuda, K.: Bayesian context clustering using cross valid prior distribution for HMM-based speech recognition, in *Proc. Interspeech'08*, 2008, 936–939.

[67] Shiota, S.; Hashimoto, K.; Nankaku, Y.; Tokuda, K.: Deterministic annealing based training algorithm for Bayesian speech recognition, in *Proc. Interspeech' 09*, 2009, 680–683.

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

[69] Ding, N.; Ou, Z.: Variational nonparametric Bayesian hidden Markov model, in *Proc. ICASSP'10*, 2010, 2098–2101.

[70] Ishiguro, K.; Yamada, T.; Araki, S.; Nakatani, T.; Sawada, H.: Probabilistic speaker diarization with bag-of-words representations of speaker angle information. IEEE Trans. Audio Speech Lang. Process., **20**, (2), (2012) 447–460.

[71] Tawara, N.; Ogawa, T.; Watanabe, S.; Kobayashi, T.: Fully Bayesian inference of multi-mixture Gaussian model and its evaluation using speaker clustering, in *Proc. ICASSP'12*, 2012, 5253–5256.

[72] Hashimoto, K.; Zen, H.; Nankaku, Y.; Masuko, T.; Tokuda, K.: A Bayesian approach to HMM-based speech synthesis, in *Proc, ICASSP 2009*, 2009, 4029–4032.

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

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

[75] Goldwater, S.; Griffiths, T.: A fully Bayesian approach to unsupervised part-of-speech tagging, in *Proc. ACL'07*, 2007, 744–751.

[76] Mochihashi, D.; Yamada, T.; Ueda, N.: Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling, in *Proc. ACL-IJCNLP*, 2009, 100–108.

[77] Watanabe, S.; Mochihashi, D.; Hori, T.; Nakamura, A.: Gibbs sampling based multi-scale mixture model for speaker clustering, in *ICASSP'11*, 2011, 4524–4527.

[78] Geman, S.; Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell., **6**, (6), (1984) 721–741.

[79] Hori, T.; Araki, S.; Yoshioka, T.; Fujimoto, M.; Watanabe, S.; Oba, T.; Ogawa, A.; Otsuka, K.; Mikami, D.; Kinoshita, K.; et al.,: Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera. IEEE Trans. Audio Speech Lang. Process., **20**, (2), (2012) 499.

[80] Fiscus, J.; Ajot, J.; Garofolo, J.: The rich transcription 2007 meeting recognition evaluation. Multimodal Technol. Percept. Humans, 2009 373–389. http://www.springerlink.com/content/94w143777u0165v5/

[81] Hinton, G.; Deng, L.; Yu, D.; Dahl, G.; Mohamed, A.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.; Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Process. Mag., **29**, (6), (2012), 82–97.

[82] Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. Ann. Stat., **1**, (2) (1973) 209–230.

[83] Griffiths, T.; Ghahramani, Z.: Infinite latent feature models and the Indian buffet process. Tech. Rep., Gatsby Unit, 2005.

[84] Teh, Y.W.; Jordan, M.I.; Beal, M.J.; Blei, D.M.: Hierarchical Dirichlet processes. J. Am. Stat. Assoc., **101**, (476), (2006) 1566–1581.

[85] Blei, D.M.; Griffiths, T.L.; Jordan, M.I.: The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *J. ACM*, **57**, (2), (2010) 7.

[86] Lee, C. y.; Glass, J.: A nonparametric Bayesian approach to acoustic model discovery, in *Proc. ACL'12*, 2012.

[87] Neubig, G.; Mimura, M.; Mori, S.; Kawahara, T.: Learning a language model from continuous speech, in *Proc. Interspeech'10*, 2010, 1053–1056.

[88] Hoffman, M.; Blei, D.; Cook, P.R.: Finding latent sources in recorded music with a shift-invariant HDP, in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2009, 438–444.

[89] Yoshii, K.; Goto, M.: Infinite latent harmonic allocation: a nonparametric Bayesian approach to multipitch analysis, in *Proc. 11th Int. Conf. Music Information Retrieval (ISMIR)*, 2010, 309–314.

[90] Nakano, M.; Le Roux, J.; Kameoka, H.; Nakamura, T.; Ono, N.; Sagayama, S.: Bayesian nonparametric spectrogram modeling based on infinite factorial infinite hidden Markov model, in *2011 IEEE Workshop on, Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, 325–328.

[91] McDermott, E.; Hazen, T.J.; Le Roux, J.; Nakamura, A.; Katagiri, S.: "Discriminative training for large-vocabulary speech recognition using minimum classification error", *IEEE Trans. Audio Speech Lang. Process.*, **15**, (1), (2007) 203–223.

[92] Gales, M.; Watanabe, S.; Fossler-Lussier, E.: Structured discriminative models for speech recognition. *IEEE Signal Process. Mag.*, **29**, (6), (2012), 70–81.

[93] Jebara, T.: Machine Learning: Discriminative and Generative, *Springer*, 2004.

Shinji Watanabe received his B.S., M.S., and Dr. Eng. degrees from Waseda University, Tokyo, Japan, in 1999, 2001, and 2006, respectively. From 2001 to 2011, he was a research scientist at NTT Communication Science Laboratories, Kyoto, Japan. From January to March in 2009, he was a visiting scholar at Georgia Institute of Technology, Atlanta, GA. Since 2011, he has been working at Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA. His research interests include Bayesian learning, pattern recognition, and speech and spoken language processing. He is a member of the Acoustical Society of Japan (ASJ) and the Institute of Electronics, Information and Communications Engineers (IEICE), and a senior member of the Institute of Electrical and Electronics Engineers (IEEE). He received the Awaya Award from the ASJ in 2003, the Paper Award from the IEICE in 2004, the Itakura Award from ASJ in 2006, and the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2006. He is currently an Associate Editor of IEEE Transactions on Audio Speech and Language Processing.

Atsushi Nakamura received the B.E., M.E., and Dr.Eng. degrees from Kyushu University, Fukuoka, Japan, in 1985, 1987, and 2001, respectively. In 1987, he joined Nippon Telegraph and Telephone Corporation (NTT), where he engaged in the research and development of network service platforms, including studies on the application of speech processing technologies to network services, at Musashino Electrical Communication Laboratories, Tokyo, Japan. From 1994 to 2000, he was

with the Advanced Telecommunications Research (ATR) Institute, Kyoto, Japan, as a Senior Researcher, undertaking research on spontaneous speech recognition, the construction of spoken language databases, and the development of speech translation systems. Since April 2000, he has been with NTT Communication Science Laboratories, Kyoto, Japan. His research interests include the acoustic modeling of speech, speech recognition and synthesis, spoken language processing systems, speech production and perception, computational phonetics and phonology, and the application of learning theories to signal analysis, and modeling.

Dr. Nakamura is a senior member of the Institute of Electrical and Electronics Engineers (IEEE), serves as a member of the IEEE Machine Learning for Signal Processing (MLSP) Technical Committee, and has served as a Vice Chair of the IEEE Signal Processing Society Kansai Chapter. He is also a member of the Institute of Electronics, Information and Communication Engineering (IEICE) and the Acoustical Society of Japan (ASJ). He received the IEICE Paper Award in 2004, and twice received the TELECOM System Technology Award of the Telecommunications Advancement Foundation, in 2006 and 2009.