SIP (2016), vol. 5, e16, page 1 of 13 © The Authors, 2016.

doi:10.1017/ATSIP.2016.15

ORIGINAL PAPER

Nested Gibbs sampling for mixture-of-mixture model and its application to speaker clustering

NAOHIRO TAWARA[1], TETSUJI OGAWA[1], SHINJI WATANABE[2] AND TETSUNORI KOBAYASHI[1]

This paper proposes a novel model estimation method, which uses nested Gibbs sampling to develop a mixture-of-mixture model to represent the distribution of the model's components with a mixture model. This model is suitable for analyzing multilevel data comprising frame-wise observations, such as videos and acoustic signals, which are composed of frame-wise observations. Deterministic procedures, such as the expectation–maximization algorithm have been employed to estimate these kinds of models, but this approach often suffers from a large bias when the amount of data is limited. To avoid this problem, we introduce a Markov chain Monte Carlo-based model estimation method. In particular, we aim to identify a suitable sampling method for the mixture-of-mixture models. Gibbs sampling is a possible approach, but this can easily lead to the local optimum problem when each component is represented by a multi-modal distribution. Thus, we propose a novel Gibbs sampling method, called "nested Gibbs sampling," which represents the lower-level (fine) data structure based on elemental mixture distributions and the higher-level (coarse) data structure based on mixture-of-mixture distributions. We applied this method to a speaker clustering problem and conducted experiments under various conditions. The results demonstrated that the proposed method outperformed conventional sampling-based, variational Bayesian, and hierarchical agglomerative methods.

Keywords: Fully Bayesian approach, Markov chain Monte Carlo, Nested Gibbs sampling, Mixture-of-mixture model, Speaker clustering

Received 17 November 2015; Accepted 28 June 2016

I. INTRODUCTION

Real-world data often comprise a set of component features, such as images made of a set of pixels and speech comprising a set of frames. These data sets have a hierarchical structure, as illustrated in Fig. 1. We describe data such as images and speech as *higher*- and *lower*-level observations. For example, in speech data obtained from a multi-party conversation, higher-level observations correspond to each speaker's utterances, where their variation is caused by the differences in the speakers. Lower-level observations correspond to frame-wise observations comprising each utterance, where their variation is caused by the differences in the contents of speech. To cluster utterances by a speaker, we need to derive a suitable mathematical representation of an utterance for extracting each speaker's characteristics independently of the contents of the their speech [1].

An effective approach for representing higher-level observations is modeling as stochastic distributions. Thus assume, we that each higher-level observation follows a unique distribution, which represents each speaker's characteristics. Members of exponential families of distributions are employed widely to model higher-level observations due to their usefulness and analytical tractability. However, the underlying assumption of uni-modality for these distributions, is sometimes too restrictive. For example, frame-wise observations, short time fast Fourier transforms of acoustic signals, and filter responses in images are known to follow multi-modal distributions, which cannot be represented by unimodal distributions [2–4]. Mixture models are reasonable approximations for representing these multi-modal distributions [5, 6] and various distributions have been used as components of mixture models such as the t-distribution [7] and von Mises–Fisher distribution [8, 9]. In particular, Gaussian distributions are used widely as a reasonable approximations for a wide class of probability distributions [10]. By using a mixture distribution to represent each cluster, the whole speaker space is modeled as a mixture of these mixture distributions. We refer to this as a *mixture-of-mixture* model. The optimal assignment of higher-level observations to clusters can be obtained by evaluating the posterior probability of assigning each observation to each cluster's mixture distribution. Thus, the clustering problem is reduced to the problem of estimating this mixture-of-mixture model.

The concept of mixture-of-mixture modeling was introduced to analyze multi-modal data sample observations

[1]Department of Communications and Computer Engineering, Waseda University, Tokyo, Japan
[2]Mitsubishi Electric Research Laboratories, MA, USA.

Corresponding author:
N. Tawara
Email: tawara@pcl.cs.waseda.ac.jp

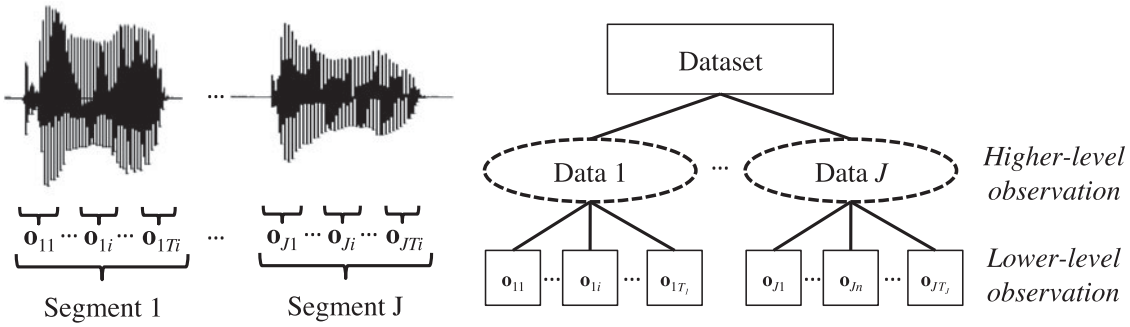


Fig. 1. Hierarchical structure of multi-level data analysis. Segment-wise (higher-level) observations are composed of a set of frame-wise (lower-level) observations. Left figure illustrates the hierarchical structure in speech data composed of frame-wise observations (e.g. Mel-frequency cepstral coefficients).

comprising both continuous and categorical variables [11, 12] and data that composed of sets of observations such as data from students nested within schools or patients within hospitals [13–15]. However, in these studies, the applications of mixture-of-mixture modeling were limited to simulated or low-dimensional data. In the present study, we focus on applying mixture-of-mixture modeling to speech data, which usually comprises high-dimensional continuous data. In [13–15], an expectation–maximization (EM) approach [16] was used to estimate mixture-of-mixture models by augmenting observations with two-level (higher-level and lower-level) latent variables. However, this maximum-likelihood-based approach often suffers from an overfitting problem when applied to high-dimensional data [1, 17]. A Bayesian approach can make the estimation of mixture-of-mixtures models more robust. For example, maximum a posterior (MAP) and variational Bayes (VB)-based methods have been applied to estimate the mixture of Gaussian mixture models (MoGMMs) [1, 18, 19]. However, the VB-based approach often still suffers from a large bias when the amount of data is limited [20]. Moreover, these methods are easily trapped by a local optimum due to the deterministic procedures in the EM-like algorithm.

To solve this problem, we propose a novel MoGMMs estimation method based on the Markov chain Monte Carlo (MCMC) method. In this approach, the optimal parameters for the MoGMMs are obtained stochastically by drawing values iteratively from their posterior distribution. These parameters can be estimated theoretically while avoiding a local optimal solution by evaluating a huge number of samples and combinations of higher-level latent variables (hLVs) Z and lower-level latent variables (lLVs) V from their joint posterior distribution $P(Z, V)$. However, in practical implementations of MCMC, evaluating such a huge number of combinations is infeasible and some approximations are required.

Previously [1, 17], we introduced a Gibbs sampling-based MoGMMs estimation approach, which draws the values of lLVs and hLVs alternately by first sampling the lLVs after initializing hLVs Z, i.e., $V \sim p(V|Z)$, and then sampling hLVs by using the fixed lLVs, i.e., $Z \sim p(Z|V)$, sampled in the previous step. This sampling method is easy to implement and highly accurate, and the it actually outperforms the VB-based approach, especially when the data are limited, e.g., when each utterance is short and the spoken utterances are few [17]. However, this sampling method, has a severe restriction because the sampling of hLVs is strictly determined by the values of the lLVs obtained in the previous sampling step. This restriction can cause the local optima problem for the hLVs, because the hLVs estimated in each iteration can be highly correlated. To solve this problem, we propose a novel sampling method for the MoGMMs based on nested Gibbs sampling, which samples both the hLVs and lLVs at the same time. This sampling method allow an enormous number of combinations of lLVs and hLVs to be evaluated efficiently, so we can find a more appropriate solution than that obtained by alternating Gibbs sampling for lLVs and hLVs.

The reminder of this paper is organized as follows. In Section II, we formulate a MoGMMs by creating a mixture-of-mixture model where each component of the mixtures is represented by a GMM. In Section III, we explain how to estimate the MoGMMs using fully Bayesian approaches based on VB and MCMC methods. In Section IV, we describe the MCMC-based model estimation method in more detail as well as the proposed nested Gibbs sampling method for MoGMMs estimation. In Section V, we present the results of speaker clustering experiments conducted to demonstrate the effectiveness of the proposed method. In Section VI, we give conclusions and discuss some directions for future research.

II. FORMULATION

In this section, we define the MoGMMs models where each component of the model is represented by a GMM. In addition, we define the generative model for segment-oriented data.

A) MoGMMs

Let $\mathbf{o}_{ut} \in \mathbb{R}^D$ be a D-dimensional observation vector, e.g., mel-frequency cepstral coefficients (MFCCs) at the t-th frame in the u-th segment, $\mathbf{O}_u \triangleq \{\mathbf{o}_{ut}\}_{t=1}^{T_u}$ is the u-th segment comprising the T_u observation vectors, and $\mathcal{O} \triangleq$

$\{\mathbf{O}_u\}_{u=1}^U$ is a set of U segments. We call this "segment-oriented data." Here, a MoGMMs is defined as follows:

$$p(\mathcal{O}|\mathbf{\Theta}) = \prod_{u=1}^{U}\sum_{i=1}^{S} h_i\, p(\mathbf{O}_u|\mathbf{\Theta}_i), \tag{1}$$

where S denotes the number of clusters; h_i represents how frequently the i-th cluster's segment appears; and $p(\mathbf{O}_u|\mathbf{\Theta}_i)$ is the likelihood of u-th segment $\mathbf{O}_u$ being assigned to the i-th cluster. In this case, $p(\mathbf{O}_u|\mathbf{\Theta}_i)$ models the intra-cluster variability for each cluster, which can be represented as:

$$p(\mathbf{O}_u|\mathbf{\Theta}_i) = \prod_{t=1}^{T_u}\sum_{j=1}^{K} w_{ij}\mathcal{N}(\mathbf{o}_{ut}|\boldsymbol{\mu}_{ij}, \mathbf{\Sigma}_{ij}), \tag{2}$$

where $\mathcal{N}$ denotes the j-th component in the i-th cluster, which is represented by a Gaussian distribution with a mean vector $\boldsymbol{\mu}_{ij}$ and a covariance matrix $\mathbf{\Sigma}_{ij}$; w_{ij}, the weight of the j-th component; and K is the number of components in each cluster's GMM. Equations (1) and (2) imply that the whole generative model for all segments $\mathcal{O}$ can be represented by a hierarchically structured MoGMMs where a GMM represents a cluster's characteristics (i.e., intra-cluster variability), and that a mixture of these GMMs can represent the entire cluster space (i.e., inter-cluster variability).

To represent this hierarchical model, we introduce two types of latent variables: $\mathcal{Z} = \{z_u\}_{u=1}^U$ represents segment-level latent variables (sLVs), each of which identifies a MoGMMs component (i.e., speaker GMM) to which the u-th segment is assigned, and $\mathcal{V} = \{\mathcal{V}_u = \{v_{ut}\}_{t=1}^{T_u}\}_{u=1}^U$, represents the frame-level latent variables (fLVs), each of which identifies an intra-cluster GMM component (the cluster distribution to which the u-th segment is assigned), to which the t-th frame-wise observation in the u-th segment is assigned. For instance, the sLVs and fLVs in MoGMMs correspond to the document-level and word-level latent variables in the latent Dirichlet allocation, where discrete data are used [21]. By contrast, we focus on modeling a continuous data space with a MoGMMs in this study.

By introducing these latent variables, we can describe the conditional distributions of the observed segments given the latent variables as follows[1]:

$$p(\mathcal{O}|\mathcal{Z},\mathcal{V},\mathbf{\Theta}) = \prod_{u=1}^{U} h_{z_u}\prod_{t=1}^{T_u} w_{z_u v_{ut}}\mathcal{N}(\mathbf{o}_{ut}|\boldsymbol{\mu}_{z_u v_{ut}}, \mathbf{\Sigma}_{z_u v_{ut}}), \tag{3}$$

where $\mathbf{\Theta} \triangleq \{\{h_i\},\{w_{ij}\},\{\boldsymbol{\mu}_{ij}\},\{\mathbf{\Sigma}_{ij}\}\}$ denote the weight of the i-th intra-cluster GMM, weight, mean vector, and covariance matrix of the j-th component of the i-th intra-cluster GMM, respectively. Note that we have assumed $\mathbf{\Sigma}_{ij}$ is a diagonal covariance matrix where the (d,d)-th element is represented by $\sigma_{ij,d}$.

[1]We use the notation $p(\cdot)$ to represent continuous probability functions and $P(\cdot)$ to represent discrete probability functions.

We describe the distribution of the latent variables as follows:

$$P(\mathcal{V}|\mathcal{Z},\mathbf{w}) = \prod_{u=1}^{U}\prod_{t=1}^{T_u}\prod_{i=1}^{S}\prod_{j=1}^{K} w_{ij}^{\delta(v_{ut},j)\delta(z_u,i)}, \tag{4}$$

$$P(\mathcal{Z}|\mathbf{h}) = \prod_{u=1}^{U}\prod_{i=1}^{S} h_i^{\delta(z_u,i)}, \tag{5}$$

where $\delta(a,b)$ denotes Kronecker's delta, which takes a value of one if $a = b$, and zero otherwise.

B) Generative process and graphical model

Using a Bayesian approach, the conjugate prior distributions of the parameters are often introduced as follows:

$$p(\mathbf{\Theta}|\mathbf{\Theta}^0) = \begin{cases} \mathbf{h} \sim \mathcal{D}(\mathbf{h}^0), \\ \mathbf{w}_i \sim \mathcal{D}(\mathbf{w}^0), \\ \{\mu_{ij,d}, \Sigma_{ij,d}\} \sim \mathcal{NG}(\xi^0, \eta^0, \mu_{j,d}^0, \sigma_{j,d}^0), \end{cases} \tag{6}$$

where $\mathcal{D}(\mathbf{h}^0)$ and $\mathcal{D}(\mathbf{w}^0)$ denote Dirichlet distributions with hyper-parameters $\mathbf{h}^0$ and $\mathbf{w}^0$, respectively. $\mathcal{NG}(\xi^0, \eta^0, \mu_{j,d}^0, \sigma_{j,d}^0)$ denotes the normal inverse gamma distribution with hyper-parameters ξ^0, η^0, $\mu_{j,d}^0$, and $\sigma_{j,d}^0$.

Based on these likelihoods and prior distributions, the generative process for our model is described as follows:

(i) Initialize $\{\mathbf{h}^0, \mathbf{\Theta}^0\}$,
(ii) Draw $\mathbf{h}$ from $\mathcal{D}(\mathbf{h}^0)$,
(iii) For each segment-level mixture component (i.e., cluster) $i = 1,\ldots,S$,
  (a) Draw $\mathbf{w}_i$ from $\mathcal{D}(\mathbf{w}^0)$,
  (b) For each frame-level mixture component (i.e., inner-cluster GMM component) $j = 1,\ldots,K$,
    (1) Draw $\{\mu_{i,d}, \sigma_{i,d}\}$ from $\mathcal{NG}(\xi_j^0, \eta_j^0, \mu_{j,d}^0, \sigma_{j,d}^0)$ for each dimension $d = 1,\ldots,D$.
(iv) For each segment $u = 1,\ldots,U$,
  (a) Draw z_u from multinomial distribution $\mathcal{M}(\mathbf{h})$,
  (b) For each frame $t = 1,\ldots,T_u$,
    (1) Draw v_{ut} from $\mathcal{M}(\mathbf{w}_{z_u})$,
    (2) Draw $\mathbf{o}_{ut}$ from $\mathcal{N}(\boldsymbol{\mu}_{z_u v_{ut}}, \mathbf{\Sigma}_{z_u v_{ut}})$.

Figure 2 shows a graphical representation of this model.

III. MODEL INFERENCE BASED ON FULLY BAYESIAN APPROACH

When we use a Bayesian approach for estimating the MoGMMs, the main task is calculating posterior distributions for the latent variables $\{\mathcal{V},\mathcal{Z}\}$ and model parameter $\mathbf{\Theta}$ given observation $\mathcal{O}$:

$$p(\mathcal{V},\mathcal{Z},\mathbf{\Theta}|\mathcal{O}) = \frac{1}{H_0}\, p(\mathcal{O},\mathcal{V},\mathcal{Z},\mathbf{\Theta}). \tag{7}$$

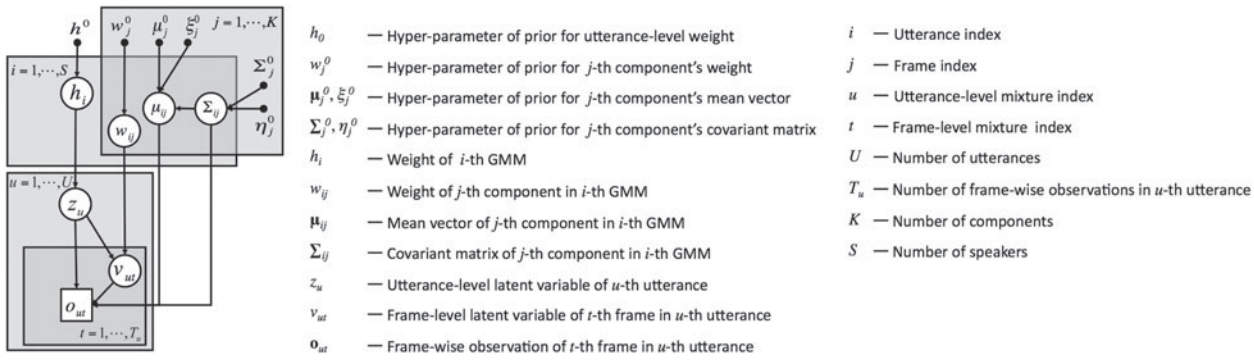


Fig. 2. Graphical representation of mixture-of-mixture model. The white square denotes frame-wise observations, and dots denote the hyper-parameters of prior distributions.

H_0 is a normalization coefficient, which is defined as follows:

$$H_0 \triangleq p(\mathcal{O}) = \sum_{\mathcal{V},\mathcal{Z}} \int p(\mathcal{O}, \mathcal{V}, \mathcal{Z}, \boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta}. \tag{8}$$

Note that the model-based clustering problem is reduced to the problem of estimating the optimal values of the fLVs and sLVs, $\{\mathcal{V}, \mathcal{Z}\}$, based on the posterior distribution defined in equation (7). Thus, the posterior probabilities of the latent variables $\mathcal{V}$ and $\mathcal{Z}$ can be calculated as follows:

$$\gamma_{v_{ut}=j|z_u=i;\boldsymbol{\Theta}} \triangleq p(v_{ut} = j|\mathbf{O}, \boldsymbol{\Theta}, z_u = i), \tag{9}$$

$$\gamma_{z_u=i;\boldsymbol{\Theta}} \triangleq p(z_u = i|\mathbf{O}, \boldsymbol{\Theta}). \tag{10}$$

Sufficient statistics of this model are computed using the aforementioned posterior probabilities as follows:

$$\begin{cases} c_i = \sum_u \gamma_{z_u=i}, \\ n_{ij} = \sum_{u,t} \gamma_{v_{ut}=j|z_u=i} \cdot \gamma_{z_u=i}, \\ \mathbf{m}_{ij} = \sum_{u,t} \gamma_{v_{ut}=j|z_u=i} \cdot \gamma_{z_u=i} \cdot \mathbf{o}_{ut}, \\ r_{ij,d} = \sum_{u,t} \gamma_{v_{ut}=j|z_u=i} \cdot \gamma_{z_u=i} \cdot (o_{ut,d})^2, \end{cases} \tag{11}$$

where c_i denotes the number of segments assigned to the i-th component of the entire cluster MoGMMs; n_{ij} is the number of frames assigned to the j-th component of the i-th intra-cluster GMM of the MoGMMs; and m_{ij} and r_{ij} are the first and second sufficient statistics, respectively.

However, it is generally infeasible to analytically estimate these posterior distributions, so we must introduce some approximations. In the rest of this section, we discuss how to approximate the posterior distributions using VB- and MCMC-based approaches.

A) Model estimation using a VB-based approach

When the VB-based model estimation method is used, the sLVs, fLVs, and model parameters are obtained deterministically by estimating their variational posterior distributions. To optimize a variational posterior distribution, we attempt to maximize the marginalized likelihood, which is described by equation (8). $p(\mathcal{V}, \mathcal{Z}, \boldsymbol{\Theta}|\mathcal{O}) = q(\mathcal{V}, \mathcal{Z})q(\boldsymbol{\Theta})$, where the optimal variational posterior distribution (i.e., the $q(\mathcal{V}, \mathcal{Z}, \boldsymbol{\Theta})$ that maximizes the free energy) can be determined as follows:

$$q(\mathcal{V}, \mathcal{Z}) \propto \exp\left(\left\langle \log p(\mathcal{O}, \mathcal{V}, \mathcal{Z}, \boldsymbol{\Theta}) \right\rangle_{q(\boldsymbol{\Theta})}\right), \tag{12}$$

$$q(\boldsymbol{\Theta}) \propto \exp\left(\left\langle \log p(\mathcal{O}, \mathcal{V}, \mathcal{Z}, \boldsymbol{\Theta}) \right\rangle_{q(\mathcal{V},\mathcal{Z})}\right). \tag{13}$$

where $\langle A \rangle_B$ denotes the expectation of A with respect to B. The optimal values of $q(\mathcal{V}, \mathcal{Z})$, and $q(\boldsymbol{\Theta})$ from equations (12) and (13) are obtained according to Algorithm 1, as follows. The posterior probability of an fLV is estimated as follows:

$$\begin{aligned} &\gamma^*_{v_{ut}=j|z_u=i;\tilde{\boldsymbol{\Theta}}} \\ &\triangleq \exp\Bigg(\langle \log w_{ij} \rangle_{q(w_{ij})} + \frac{1}{2}\sum\nolimits_d \langle \log \sigma_{ij,d} \rangle_{q(\sigma_{ij,d})} \\ &\quad - \frac{D}{2}\log 2\pi - \frac{1}{2}\sum\nolimits_d \left\langle \frac{(o_{ut,d} - \mu_{ij,d})^2}{\sigma_{ij,d}} \right\rangle_{q(\mu_{ij,d}|\sigma_{ij,d})}\Bigg). \end{aligned} \tag{14}$$

We can determine the posterior distribution of an fLV by normalizing (14) as follows:

$$q(v_{ut} = j|z_u = i) = \frac{\gamma^*_{v_{ut}=j|z_u=i;\tilde{\boldsymbol{\Theta}}}}{\sum_j \gamma^*_{v_{ut}=j|z_u=i;\tilde{\boldsymbol{\Theta}}}}. \tag{15}$$

In the same manner, we can compute an sLV $\gamma_{z_u=i;\tilde{\boldsymbol{\Theta}}}$ from the posterior probability $\gamma^*_{z_u=i;\tilde{\boldsymbol{\Theta}}}$ as follows:

$$\gamma^*_{z_u=i;\tilde{\boldsymbol{\Theta}}} \triangleq \exp\left(\langle \log h_i \rangle_{q(h_i)} + \sum_t \log \sum_j \gamma^*_{v_{ut}=j|z_u=i;\tilde{\boldsymbol{\Theta}}}\right), \tag{16}$$

$$q(z_u = i) = \frac{\gamma^*_{z_u=i;\tilde{\boldsymbol{\Theta}}}}{\sum_i \gamma^*_{z_u=i;\tilde{\boldsymbol{\Theta}}}}. \tag{17}$$

The expected values of the parameters described in (14)–(16) are computed as follows:

$$\langle \log h_i \rangle_{q(h_i)} = \psi(\tilde{h}_i) - \psi\left(\sum_i \tilde{h}_i\right), \tag{18}$$

$$\langle \log w_{ij} \rangle_{q(w_{ij})} = \psi(\tilde{w}_{ij}) - \psi\left(\sum_j \tilde{w}_{ij}\right), \tag{19}$$

$$\langle \log \sigma_{ij,d} \rangle_{q(\sigma_{ij,d})} = \psi(\tilde{\eta_{ij}}) - \log \tilde{\sigma}_{ij,d}, \tag{20}$$

$$\left\langle \frac{(o_{ut,d} - \mu_{ij,d})^2}{\sigma_{ij,d}} \right\rangle_{q(\mu_{ij,d}|\sigma_{ij,d})} = \frac{\tilde{\eta}_{ij}(o_{ut,d} - \tilde{\mu}_{ij,d})^2 + \tilde{\xi}_{ij}}{\tilde{\sigma}_{ij,d}}, \tag{21}$$

where $\psi(\cdot)$ denotes the digamma function and $\tilde{\boldsymbol{\Theta}} = \{\tilde{h}_i, \tilde{w}_{ij}, \tilde{\xi}_{ij}, \tilde{\eta}_{ij}, \tilde{\boldsymbol{\mu}}_{ij}\}$ are the hyper-parameters of the posterior distributions for $\tilde{\boldsymbol{\Theta}}$, which are computed as follows:

$$\tilde{\boldsymbol{\Theta}} = \begin{cases} \tilde{h}_i = h^0 + c_i, \\ \tilde{w}_{ij} = w_j^0 + n_{ij}, \\ \tilde{\xi}_{ij} = \xi^0 + n_{ij}, \\ \tilde{\eta}_{ij} = \eta^0 + n_{ij}, \\ \tilde{\boldsymbol{\mu}}_{ij} = \tilde{\xi}_{ij}^{-1}(\xi^0 \boldsymbol{\mu}_j^0 + \mathbf{m}_{ij}), \\ \tilde{\sigma}_{ij,d} = \sigma_{j,d}^0 + r_{ij,d} + \xi^0(\mu_{j,d}^0)^2 - \tilde{\xi}_{ij}(\tilde{\mu}_{ij,d})^2. \end{cases} \tag{22}$$

Algorithm 1 shows the VB-based model estimation algorithm. The fLVs and sLVs that maximize (equations (15) and (17)) are the MAP values of their posterior distributions, where we assume that these MAP values are the optimal clustering results.

Algorithm 1: Model estimation algorithm using the VB method.

```
initialize Θ̃;
repeat
    for all clusters i and components j do
        for all segments u and frames t do
            Compute γ_q(V, Z) in equation (12) before
            computing the expectation values
            described in equations (15) and (17);
    for all clusters i and components j do
        Compute the hyper-parameters of q(Θ) in
        equation (13) using the sufficient statistics, as
        described in equations (18)–(21)
until converged ;
```

This VB-based procedure monotonically increases the free energy, as described in equation (8) under the variational posterior distribution $q(\mathcal{V}, \mathcal{Z}, \boldsymbol{\Theta})$, but this approach suffers from two problems, which are caused by the difference between true and variational posterior distributions, as well as the biased values. The first problem is that the true posterior distributions of fLVs, sLVs, and the model parameters in MoGMMs cannot be factorized (i.e., $p(\boldsymbol{\Theta}, \mathcal{V}, \mathcal{Z}|\mathcal{O}) \neq p(\mathcal{V}|\mathcal{Z}, \mathcal{O})\, p(\mathcal{Z}|\mathcal{O})\, p(\boldsymbol{\Theta}|\mathcal{O})$), although the variational posterior distributions assume that they can. The second problem is that the posterior probability obtained is generally biased because the calculated statistics are strongly biased by the size of each segment. These problems are especially severe when the number of segments is limited. To solve these problems, we need to estimate the marginalized posterior distributions, into which model parameter $\boldsymbol{\Theta}$ and fLVs $\mathcal{V}$ are collapsed [2]. This is obtained by marginalizing equation (7) with respect to these parameters as follows:

$$P(\mathcal{V}, \mathcal{Z}|\mathcal{O}) = \frac{1}{H_0} \int p(\mathcal{V}, \mathcal{Z}, \boldsymbol{\Theta}, \mathcal{O}).\mathrm{d}\boldsymbol{\Theta}. \tag{23}$$

We can then estimate the posterior distribution of the latent variables directly and obtain an unbiased estimation.

Collapsed VB methods for estimating the marginalized posterior distribution have been proposed in several studies [22, 23], but these approaches are generally infeasible for our hierarchical model because we cannot apply the approximation of convexity to a hierarchical structure. Therefore, we introduce the MCMC method to estimate the marginalized posterior distribution from equation (23).

B) Model estimation based on the MCMC approach

Using an MCMC-based approach, we obtain samples of latent variables directly from their posterior distributions. We can derive the marginalized distribution with respect to the model parameters described in equation (23) because we do not need to evaluate the normalization term equation (8) when employing an MCMC approach.

1) Marginalized likelihood for complete data

First, we derive the logarithmic marginalized likelihood for the complete data, $\log p(\mathcal{O}, \mathcal{V}, \mathcal{Z})$. In the case of complete data, we can utilize all the alignments of observations $\mathbf{o}_{ut}$ to a specific Gaussian component distribution because all of the latent variables, $\{\mathcal{V}, \mathcal{Z}\}$, are treated as observations. Then, the posterior distributions for each of the latent variables, $P(z_u = i|\cdot)$ and $P(v_{ut} = j|\cdot)$ for all i, j, u, and t, return 0 or 1 based on the assigned information. Thus, $\gamma_{v_{ut}=j|z_u=i}$ and $\gamma_{z_u=i}$ described by equations (9) and (10) are zero-or-one values depending on the assignment of the data. Then, the sufficient statistics of this model, then, can be represented as follows:

$$\begin{cases} c_i = \sum_u \delta(z_u, i), \\ n_{ij} = \sum_{u,t} \delta(z_u, i) \cdot \delta(v_{ut}, j), \\ \mathbf{m}_{ij} = \sum_{u,t} \delta(z_u, i) \cdot \delta(v_{ut}, j) \cdot \mathbf{o}_{ut}, \\ r_{ij,d} = \sum_{u,t} \delta(z_u, i) \cdot \delta(v_{ut}, j) \cdot (o_{ut,d})^2, \end{cases} \tag{24}$$

[2] In this case, "collapsed" means that samples are drawn from the marginalized distribution with respect to the model parameter $\boldsymbol{\Theta}$. In the following, we refer to collapsed Gibbs sampling simply as Gibbs sampling.

We can analytically derive the logarithmic marginalized likelihood for the complete data by substituting equations (3)–(5) into the following integration equation:

$$\begin{aligned}
&\log p(\mathcal{V},\mathcal{Z},\mathcal{O}) \\
&= \log \int p(\mathcal{V},\mathcal{Z},\mathcal{O}|\boldsymbol{\Theta})\, p(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta} \\
&= \log \frac{\Gamma(h^0)\prod_i \Gamma(\tilde{h}_i)}{\Gamma(h^0)^S \Gamma\left(\sum_i \tilde{h}_i\right)} + \log \prod_i \frac{\Gamma\left(\sum_j w_j^0\right)\prod_j \Gamma(\tilde{w}_{ij})}{\prod_j \Gamma(w_j^0)\Gamma\left(\sum_j \tilde{w}_{ij}\right)} \\
&\quad + \beta \log \prod_{i,j} (2\pi)^{-\frac{n_{ij} D}{2}} \frac{(\xi^0)^{\frac{D}{2}} \left(\Gamma\left(\frac{\eta_j^0}{2}\right)\right)^{-D} \left(\prod_d \sigma_{j,dd}^0\right)^{\frac{\eta_j^0}{2}}}{(\tilde{\xi}_{ij})^{\frac{D}{2}} \left(\Gamma\left(\frac{\tilde{\eta}_{ij}}{2}\right)\right)^{-D} \left(\prod_d \tilde{\sigma}_{ij,dd}\right)^{\frac{\tilde{\eta}_{ij}}{2}}},
\end{aligned} \tag{25}$$

where $\tilde{\boldsymbol{\Theta}}_{ij} \triangleq \{\tilde{h}_i, \tilde{w}_{ij}, \tilde{\xi}_{ij}, \tilde{\eta}_{ij}, \tilde{\mu}_{ij,d}, \tilde{\sigma}_{ij,d}\}$ denotes the hyperparameter of the marginalized likelihood defined in equation (22).

To construct the MCMC sampler, we define the following logarithmic likelihood function for the complete data using simulated annealing (SA) [24]:

$$\begin{aligned}
H_p(\beta) &\triangleq \log p_\beta(\mathcal{V},\mathcal{Z},\mathcal{O}) \\
&= \log p(\mathcal{O}|\mathcal{V},\mathcal{Z}) + \frac{1}{\beta}\log P(\mathcal{V},\mathcal{Z}),
\end{aligned} \tag{26}$$

where β is an inverse temperature defined for SA, which controls the speed of convergence. We can now derive the posterior distribution as follows:

$$\begin{aligned}
P(\mathcal{V},\mathcal{Z}|\mathcal{O}) &= \frac{1}{H_p(\beta)} p(\mathcal{V},\mathcal{Z})\, p(\mathcal{O}|\mathcal{V},\mathcal{Z})^\beta \\
&= \frac{1}{H_p(\beta)} \exp\{-\beta H(\boldsymbol{\Psi})\},
\end{aligned} \tag{27}$$

where $H_p(\beta)$ is a normalization term introduced to normalize $\{\mathcal{V},\mathcal{Z}\}$ under the temperature β. The goal of the MCMC approach is to obtain samples from equation (27). In the next section, we discuss how to design the sampler in order to obtain samples from this posterior distribution.

IV. IMPLEMENTATION OF MCMC-BASED MODEL ESTIMATION

We introduce a collapsed Gibbs sampler [25] to obtain samples of sLVs and fLVs from their posterior distributions. Previously, we introduced a Gibbs assumption that alternates the sampling of fLVs with some initializations of sLVs, before sampling the sLVs using the fixed fLVs sampled in the previous step [1, 17]. The drawback of this approach is that the sampling of sLVs is determined strictly by the values of the fLVs obtained in the previous sampling step and the sLVs estimated in each iteration can be highly correlated. To solve this problem, we propose a novel sampling method that samples both sLVs and fLVs at the same time. This sampling method allows an enormous number of combinations of fLVs and sLVs to be evaluated efficiently, so we can find a more appropriate solution than that obtained when alternating Gibbs sampling for fLVs and sLVs. We refer to this novel sampling technique as nested Gibbs sampling. This section describes its formulation and implementation.

A) Nested Gibbs sampling for MoGMMs

For Gibbs sampling, we draw the value of each variable iteratively from its posterior distributions and conditioning it with the sampled values of the other variables. This posterior distribution is called the "proposal distribution." In the case of MoGMMs, the proposal distribution is the joint posterior distribution of the sLV and fLVs related to the u-th segment of $\{\mathcal{V}_u, z_u\}$, which is conditioned on the sampled value of the latent variables related to the other segments $\{\mathcal{V}^*_{\backslash u}, \mathcal{Z}^*_{\backslash u}\}$. Therefore, the proposal distribution of MoGMMs is described as follows:

$$P\left(\mathcal{V}_u, z_u | \left\{\mathcal{V}^*_{\backslash u}, \mathcal{Z}^*_{\backslash u}\right\}, \mathcal{O}\right) = \frac{p\left(\mathcal{V}_u, z_u, \left\{\mathcal{V}^*_{\backslash u}, \mathcal{Z}^*_{\backslash u}\right\}, \mathcal{O}\right)}{p\left(\left\{\mathcal{V}^*_{\backslash u}, \mathcal{Z}^*_{\backslash u}\right\}, \mathcal{O}\right)}, \tag{28}$$

where $\mathcal{V}^*_{\backslash u} = \{v^*_{u't} | \forall u' \neq u, \forall t\}$ and $\mathcal{Z}^*_{\backslash u} = \{z^*_{u'} | \forall u' \neq u\}$ denote the sets of samples for fLVs and sLVs, respectively, except for those related to the u-th segment. After some iterative sampling using equation (28), the samples obtained are approximately distributed according to their true posterior distributions. Direct sampling from a proposal distribution equation (28) is theoretically feasible because equation (28) takes the form of a multinomial distribution. However, it is impractical to evaluate an enormous number of possible combinations of solutions. We notice that it is enough to estimate the value of sLVs in order to estimate the optimal assignment of utterances to speaker clusters. Therefore, we try to marginalize out fLVs in equation (28) to make the computation simple. We propose an MCMC-based approach, which samples the value of z_u directly from the following marginalized posterior instead of equation (28):

$$\begin{aligned}
p\left(z_u | \left\{\mathcal{V}^*_{\backslash u}, \mathcal{Z}^*_{\backslash u}\right\}, \mathcal{O}\right) = \int & p\left(z_u | \mathcal{V}^*_u, \left\{\mathcal{V}^*_{\backslash u}, \mathcal{Z}^*_{\backslash u}\right\}, \mathcal{O}\right) \\
& \times p\left(\mathcal{V}^*_u | \left\{\mathcal{V}^*_{\backslash u}, \mathcal{Z}^*_{\backslash u}\right\}, \mathcal{O}\right) \mathrm{d}\mathcal{V}_u.
\end{aligned} \tag{29}$$

However, this integration is also infeasible because each v_{ut} in $\mathcal{V}_u$ takes one of the number of K values (i.e., the number of GMM components) and their combination are exponentially large. Therefore, we introduce an approximated approach, which uses the sampled value of $\mathcal{V}^{**}_u$ obtained from its true posterior $p(\mathcal{V}^*_u | \{\mathcal{V}^*_{\backslash u}, \mathcal{Z}^*_{\backslash u}\}, \mathcal{O})$. Then, $\mathcal{V}^*_u$ is marginalized out from equation (28) using the sampled value $\mathcal{V}^{**}_u$ by the following approximation:

$$p\left(z_u | \left\{\mathcal{V}^*_{\backslash u}, \mathcal{Z}^*_{\backslash u}\right\}, \mathcal{O}\right) \simeq \sum_{\mathcal{V}^{**}_u} P\left(z_u | \mathcal{V}^{**}_u, \left\{\mathcal{V}^*_{\backslash u}, \mathcal{Z}^*_{\backslash u}\right\}, \mathcal{O}\right). \tag{30}$$

We can easily sample the value of z_u from equation (30) because this is a multinomial distribution over z_u that takes the one of the C (i.e., number of clusters) values. With this approach, the Gibbs sampling chain for z_u is followed by another Gibbs sampling chain in which we sample the values of $\mathcal{V}_u$ from its posterior distribution, conditioned on any potential value of z_u. We refer to this Gibbs sampler for $\mathcal{V}_u$ as a sub-Gibbs sampler and we refer to the obtained samples as $\mathcal{V}^{**}_{u|z_u=i} = \{v^{**}_{ut|z_{u=i}}\}_{t=1}^{T_u}$. In the sub-Gibbs sampler, each value of $v^{**}_{ut|z_u=i}$ is sampled for all i as follows:

$$v^{**}_{ut|z_u=i} \sim P\left(v_{ut}|\left\{\mathcal{V}^*_{\backslash u}, \mathcal{Z}^*_{\backslash u}\right\}, \mathcal{V}^{**}_{u\backslash t}, z_u = i, \mathcal{O}\right), \quad (31)$$

where $\mathcal{V}^{**}_{u\backslash t|z_u=i} = \{v^{**}_{ut'|z_u=i}|\forall t' \neq t\}$ denotes the samples of fLVs obtained from the sub-Gibbs sampler that are related to all of the frames, except to the t-th frame in the u-th segment. After several iterations of equation (31) for all t in the u-th segment, we obtain N^{Gibbs} samples. We then draw a sample of sLV for the u-th segment from its posterior distribution conditioned on the samples $\{\mathcal{V}^{**}_{u|z_u=i}{}^{(n)}\}_{n=1}^{N^{Gibbs}}$. By aggregating the value of $\{\mathcal{V}^{**}_{u|z_u=i}{}^{(n)}\}_{n=1}^{N^{Gibbs}}$ over all possible values of i, the Gibbs sampler for z_u is defined as follows:

$$\begin{aligned} z_u &\sim P(z_u|\left\{\mathcal{V}^*_{\backslash u}, \mathcal{Z}^*_{\backslash u}\right\}, \mathcal{O}) \\ &= \sum_{\forall \mathcal{V}_u} P\left(\mathcal{V}_u, z_u|\left\{\mathcal{V}^*_{\backslash u}, \mathcal{Z}^*_{\backslash u}\right\}, \mathcal{O}\right) \\ &= \sum_{\forall \mathcal{V}_u} P\left(z_u|\left\{\mathcal{V}^*_{\backslash u}, \mathcal{Z}^*_{\backslash u}\right\}, \mathcal{V}_u, \mathcal{O}\right) P\left(\mathcal{V}_u|\left\{\mathcal{V}^*_{\backslash u}, \mathcal{Z}^*_{\backslash u}\right\}, \mathcal{O}\right). \end{aligned} \quad (32)$$

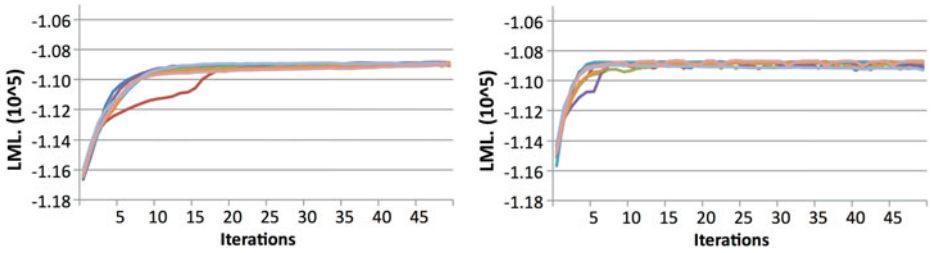
By aggregating N^{Gibbs} samples of $\mathcal{V}_u$ from $p(\mathcal{V}_u|z_u = i, \{\mathcal{V}^*_{\backslash u}, \mathcal{Z}^*_{\backslash u}\}, \mathcal{O})$ for all possible values of i and then plugging them into $p(z_u|V_u, O)$, we obtain the following Monte Carlo integration:

$$\begin{aligned} &\sum_{\forall \mathcal{V}_u} P\left(z_u|\left\{\mathcal{V}^*_{\backslash u}, \mathcal{Z}^*_{\backslash u}\right\}, \mathcal{V}_u, \mathcal{O}\right) P\left(\mathcal{V}_u|\left\{\mathcal{V}^*_{\backslash u}, \mathcal{Z}^*_{\backslash u}\right\}, \mathcal{O}\right) \\ &\simeq \frac{1}{N^{Gibbs}} \sum_{n=1}^{N^{Gibbs}} P\left(z_u|\mathcal{V}^{**(n)}_u, \left\{\mathcal{V}^*_{\backslash u}, \mathcal{Z}^*_{\backslash u}\right\}, \mathcal{O}\right). \end{aligned} \quad (33)$$

We refer to these procedures as *nested Gibbs sampling*, because we sample z_u from equation (33) using the value of $\mathcal{V}^{**(n)}_u$ which can be obtained from the sub-Gibbs sampler defined by equation (31) in a nested manner. A large number of samples, N^{Gibbs}, may be required to accurately represent of the marginal value for equation (33). To evaluate the effect of the number of samples on the overall sampling procedure, we applied the proposed nested Gibbs sampler to practical speech data. Figure 3 shows the logarithmic marginalized likelihoods (LMLs) obtained using the proposed nested Gibbs sampling method with different sampling sizes. The eight lines in each figure correspond to the results of eight trials with different random seeds. This figure shows that high accuracy may be achieved with a small number of samples, and that even one sample may be adequate to approximate the marginal value in equation (33). Algorithm 2 shows the algorithm of the

Algorithm 2: Model estimation algorithm based on the proposed nested Gibbs sampling method.

```
initialization {V**, Z**}, V*;
repeat
    for all segments u do
        for all clusters i do
            for all frames t do
                for all components j do
                    Update γ^β_{v_ut=j|z_u=i} ←
                    P_β(v_ut = j | {Z*_\u, V*_\u} V**_{u\t}; z_u = i);
                Draw the values of the fLVs, v**_ut, from
                their posterior probability with
                v**_ut ~ γ^β_{v_ut=·|z_u=i};
            Update
            γ^β_{z_u=i|V*_u} ← P_β(z_u = i | {Z*_\u, V*_\u}, V**_u);
        Draw the value of the sLVs, z*_u, from their
        posterior distribution with z*_u ~ γ^β_{z_u=i|V**_u};
        Update the values of the fLVs with V*_u ← V**_u;
        Update the SA temperature β with respect to
        scheduling
until some conditions are met ;
```



$N^{Gibbs} = 1$ (a) $N^{Gibbs} = 5$ (b)

Fig. 3. LML obtained using proposed nested Gibbs sampler, applied to $A1$ + station noise. Refer to Table 1 for the details of test set A1. Each figure shows results with a different sampling size N^{samp}. Eight lines correspond to results of eight trials using different random seeds. (a) $N^{Gibbs} = 1$ & (b) $N^{Gibbs} = 5$

nested Gibbs sampler for MoGMMs. The formulations of equations (31) and (33) are described in the Appendix.

B) Computation of the marginalized likelihood

For the Gibbs sampler described in A, we can approximate the joint likelihood equation (26) using the sampled latent variables $\{\mathcal{V}_u^*, \mathcal{Z}_u^*\}_{u=1}^U$.

Figure 4 is a scatter diagram showing the marginalized likelihood and K values (which are used widely for the measurement of the clustering) calculated from the results obtained when the proposed nested Gibbs sampler was applied to B1 and B1 with four types of noise. The values of K are explained in the Experiment section. The differences in the plots indicate the distinct speakers. This figure shows that the value of K is strongly correlated with the marginalized likelihood. Therefore, we can use the marginalized likelihood as a measure of the appropriateness of the models.

V. SPEAKER CLUSTERING EXPERIMENTS

We investigated the effectiveness of our model optimization methods at speaker clustering using the TIMIT [26] and CSJ [27] databases. We compared the following four model estimation methods:

- **n-Gibbs:** MCMC-based model estimation using the proposed nested Gibbs sampling method.
- **Gibbs:** MCMC-based model estimation using conventional Gibbs sampling where the fLVs and sLVs are sampled alternately [1, 17].
- **VB:** VB-based model estimation [19].
- **HAC-GMM:** hierarchical agglomerative clustering method. A GMM is estimated for each utterance in a maximum-likelihood manner. The similarity between utterances is defined as the cross likelihood ratio between corresponding GMMs. The pair of utterances with the greatest similarity is merged iteratively until the correct number of speakers is obtained [3].

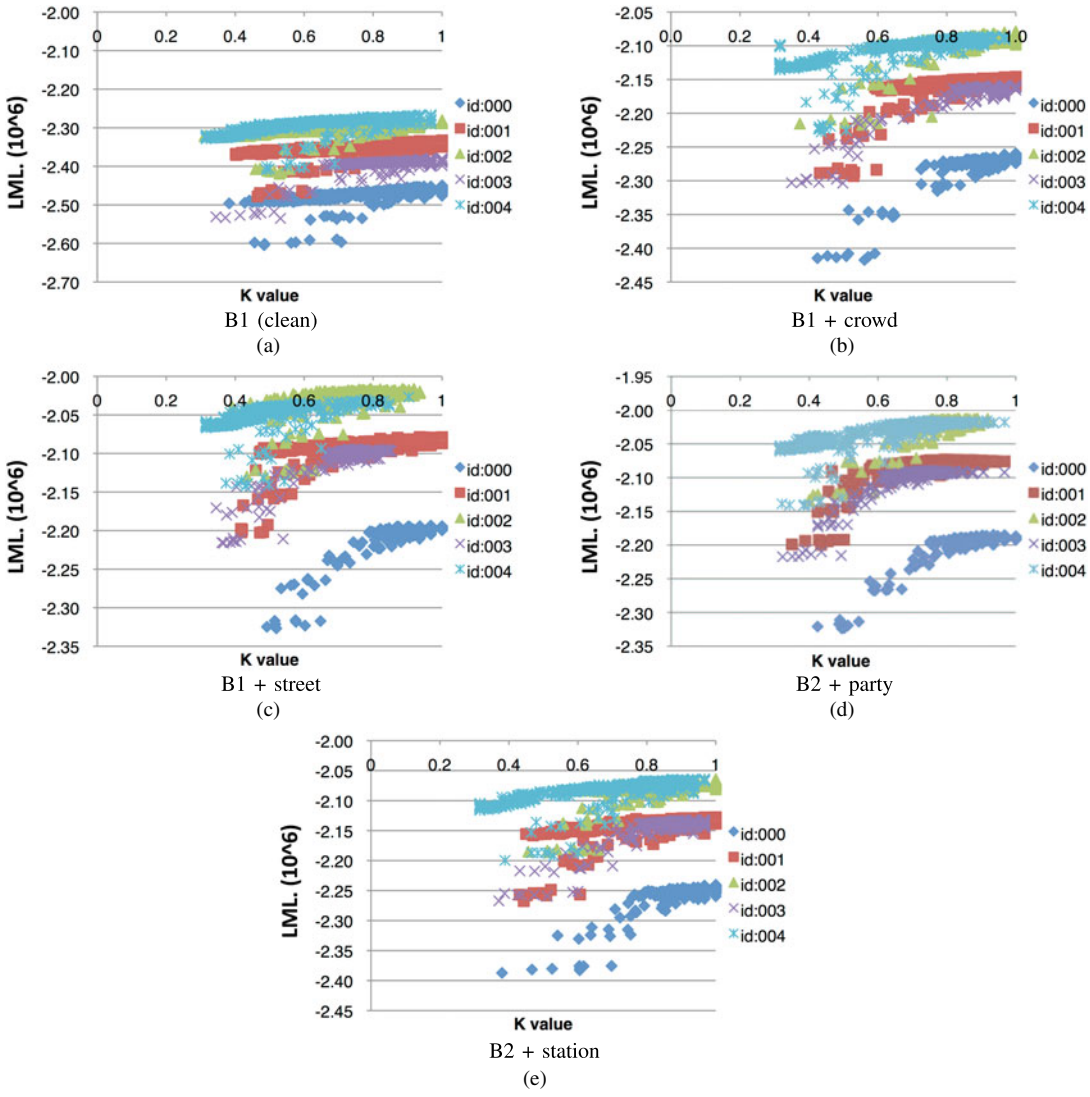


Fig. 4. LML as a function of K value. Each plot shows the results obtained by applying the proposed n-Gibbs sampler to five different datasets (id:000, 001, . . . , 004). Refer to Table 1 for the details of test set B1. (a) B1 (clean) & (b) B1 + crowd (c) B1 + street & (d) B2 + party (e) B2 + station

A) Experimental setup

1) Datasets

All of the experiments were conducted using 11 evaluation sets obtained from TIMIT and CSJ. Table 1 lists the number of speakers and utterances in the evaluation sets used. T1 and T2 were constructed using TIMIT. T1 corresponds to the core test set of TIMIT, which includes 192 utterances from 24 speakers. T2 is the complete test set, which includes 1152 utterances from 144 speakers. In this case, there were no overlaps between T1 and T2. The remaining nine evaluation sets were constructed using CSJ as follows: all lecture speech in CSJ was divided into utterance units based on the silence segments in their transcriptions, five speakers were then randomly selected, and five, 10, and 20 of their utterances were chosen for A1, A2, and A3, respectively. In the same manner, we randomly selected 10 and 15 different speakers and five, 10, and 20 of their utterances were used for B1 to B3 and C1 to C3, respectively. We evaluated five combinations of different speakers for each dataset. The resulting clustering performance for each dataset was the average of these five combinations.

The speech data from TIMIT and CSJ are not corrupted by noise. In additional experiments, we used noisy speech data, which we created by overlapping each utterance with four types of non-stationary noise (crowd, street, party, and station) selected from the noise database of the Japan Electronic Industry Development Association [28]. These noises were overlapped with each utterance at a signal-to-noise ratio of about 10 dB. Speech data were sampled at 16 kHz and quantized into 16-bit data. We used 26-dimensional acoustic feature parameters, which comprised 12-dimensional MFCCs with log energy and their Δ parameters. The frame length and frame shift were 25 and 10 ms, respectively.

2) Measurement

We employed the average cluster purity (ACP), average speaker purity (ASP), and their geometric means (K value) as the speaker clustering evaluation criteria [29]. In this experiment, the correct speaker label was manually annotated for each utterance. Let S_T be the correct number of speakers; S is the estimated number of speakers; n_{ij} is the estimated number of utterances assigned to speaker cluster i in all utterances by speaker j; n_j is the estimated number of utterances of speaker j; n_i is the estimated number of utterances assigned to speaker cluster i; and U is the number of all utterances. The cluster purity p_i and speaker purity q_j were then calculated as follows.

$$p_i = \sum_{j=0}^{S_T} \frac{n_{ij}^2}{n_i^2}, \quad q_j = \sum_{i=0}^{S} \frac{n_{ij}^2}{n_j^2}. \tag{34}$$

The cluster purity is the ratio of utterances derived from the same speaker relative to the utterances assigned to each cluster. The speaker purity is the ratio of utterances assigned to the same cluster relative to the utterances spoken by each speaker. Thus, ACP and ASP are calculated as follows.

$$V_{\mathrm{ACP}} = \frac{1}{U} \sum_{i=0}^{S} p_i n_i, \quad V_{\mathrm{ASP}} = \frac{1}{U} \sum_{j=0}^{S_T} q_j n_j. \tag{35}$$

The K value is obtained as the geometric mean between ACP and ASP as follows:

$$K = \sqrt{V_{ACP} \cdot V_{ASP}}. \tag{36}$$

Table 1. Details of test set.

Test set	Number of speakers	Number of utterances	Average total duration (min)
T1	24	192	9.7
T2	144	1152	58.8
A1	5	25	2.8
A2	5	50	5.6
A3	5	100	11.1
B1	10	50	5.6
B2	10	100	11.3
B3	10	200	22.5
C1	15	75	13.0
C2	15	150	26.0
C3	15	300	51.8

3) Evaluation conditions

The number of iterations was set to 100 in the MCMC-based method, which was sufficient for convergence in both the conventional and proposed Gibbs sampling in all of the following experimental conditions. We conducted the same speaker clustering experiment eight times using different seeds each time. We evaluated the marginalized likelihood described in equation (26) for each result and selected the result with the highest likelihood from those obtained during the 100 iterations of eight experiments.

The hyper-parameters in equation (22) were set as follows: $\mathbf{w}^{(0)} = \{\rho, \ldots, \rho\}$ for all components; $h^0 = \rho$ and $\mathbf{h}^{(0)} = \{\rho, \ldots, \rho\}$ for all clusters; $\eta^{(0)} = 1$ and $\xi^{(0)} = \rho$; $\boldsymbol{\mu}^{(0)} = \boldsymbol{\mu}(\mathcal{O})$ and $\boldsymbol{\Sigma}^{(0)} = \eta^0 \boldsymbol{\Sigma}(\mathcal{O})$, where $\boldsymbol{\mu}(\mathcal{O})$ and $\boldsymbol{\Sigma}(\mathcal{O})$ were the mean vectors and covariance matrices estimated from the whole dataset, respectively. The value range for ρ was $\{1, 10, 100, 1000\}$. These parameters were determined using the development data set obtained from the CSJ dataset. We initialized both the sLVs and fLVs randomly.

B) Experimental results

1) Comparison with the conventional Gibbs sampler

We evaluated conventional Gibbs sampling and the proposed nested Gibbs sampling method with different numbers of mixture components using both clean and noisy datasets. Figure 5 shows the K values obtained using the **Gibbs** and **n-Gibbs** samplers with different numbers of mixture components when they were applied to clean data (A1) and noisy data (A1 + crowd). We can see that the highest K value was obtained when one or two components were used for both the **Gibbs** and **n-Gibbs** samplers. This

indicates that a small number of Gaussian distributions are sufficient to model each speaker's utterances in either sampling method when clean data are used. However, in the case of noisy data, the nested Gibbs sampler performed best with eight components of mixtures, but the conventional Gibbs sampler with eight components achieved worse results than the proposed method. This suggests that samples from noisy data follow a multi-modal distribution and that the proposed sampling method can represent this multi-modality. By contrast, the conventional Gibbs sampler could not model these complex data even with a large number of mixture components. Later, we will discuss the reason why the conventional Gibbs sampler degraded the K value for the noisy data set by using diagrams to show the convergence of the samplers.

Figure 6 shows the logarithmic marginalized likelihoods of the samples obtained using the conventional Gibbs and proposed nested Gibbs sampling methods when applied to A1 with different SA temperatures [24]. The eight lines in these figures represent the results obtained from eight trials with different seeds. We can see that no trial converged to a unique distribution without SA (i.e., $\beta^{init} = 1$) when a conventional Gibbs sampler was used. Introducing a higher temperature ($\beta^{init} = 30$) offered some protection from divergence, but large variations still remained, as shown in Figs 6(c) and 6(e). These results indicate that the conventional Gibbs sampler was often trapped by a local optimum. However, in the case of the nested Gibbs sampler, the likelihoods converged after only 20 iterations at most, and all of the trials converged to almost the same result, even when we did not use the SA method (i.e., $\beta^{init} = 1$). These results indicate the greater effectiveness of the proposed sampling method.

Tables 2 and 3 list the K values obtained using each method for clean and noisy speech data, respectively. These tables demonstrate that the nested Gibbs sampler outperformed the conventional Gibbs sampler irrespective of the evaluation sets, under clean and noisy conditions. These results imply that the proposed method can model data drawn from both single and multi-modal distributions, which the conventional Gibbs sampler was unable to calculate.

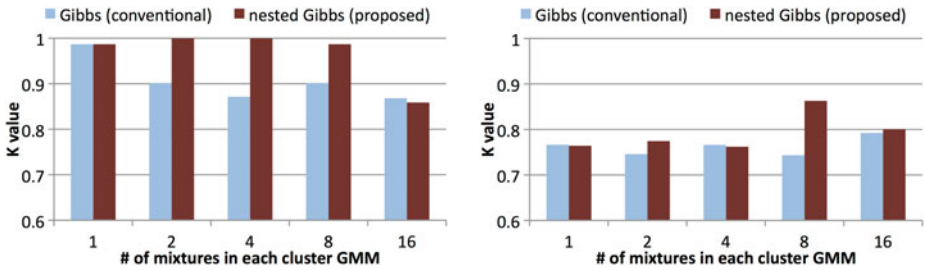


Fig. 5. K values obtained by existing Gibbs and proposed nested Gibbs sampler applied on (a) clean (A1) and (b) noisy (A1 + crowd) speech.

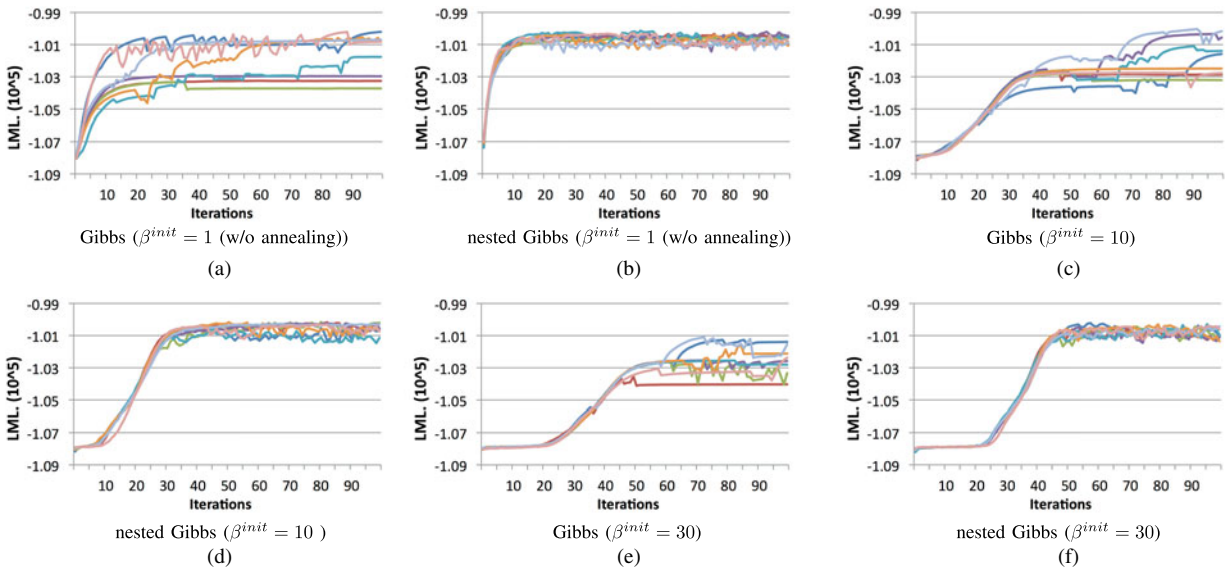


Fig. 6. LML obtained by Gibbs and nested Gibbs with SA applied on A1. Each figure shows result with different initial temperature β^{init}. Eight lines correspond to the results of eight trials with different seeds. (a) Gibbs ($\beta^{init} = 1$ (w/o annealing)); (b) nested Gibbs ($\beta^{init} = 1$ (w/o annealing)); (c) Gibbs ($\beta^{init} = 10$); (d) nested Gibbs ($\beta^{init} = 10$); (e) Gibbs ($\beta^{init} = 30$); (f) nested Gibbs ($\beta^{init} = 30$).

Table 2. K-value for clean test sets.

Evaluation data	n-Gibbs	Gibbs	VB	HAC-GMM
T1 (spkr:24 utt:192)	**0.96**	0.84	0.74	0.88
T2 (spkr:144 utt:1152)	**0.74**	0.52	0.41	0.73
A1 (spkr:5 utt:25)	**1.00**	0.90	0.92	0.93
A2 (spkr:5 utt:50)	**0.99**	0.96	0.97	**0.99**
A3 (spkr:5 utt:100)	0.98	0.97	**0.99**	0.97
B1 (spkr:10 utt:50)	**0.98**	0.93	0.85	0.95
B2 (spkr:10 utt:100)	**0.98**	0.90	0.90	0.96
B3 (spkr:10 utt:200)	**0.98**	0.91	0.96	0.96
C1 (spkr:15 utt:75)	**0.97**	0.92	0.81	0.95
C2 (spkr:15 utt:150)	0.93	0.91	0.90	**0.96**
C3 (spkr:15 utt:300)	0.92	0.91	0.91	**0.95**

Table 3. K value for noisy test sets. Four types of noise (crowd, street, party, and station) are overlapped with speech of nine datasets.

Evaluation data	n-Gibbs	Gibbs	VB	HAC-GMM
A1 + crowd (spkr:5 utt:25)	**1.00**	0.90	0.82	0.95
A2 + crowd (spkr:5 utt:50)	**0.99**	0.96	0.95	0.97
A3 + crowd (spkr:5 utt:100)	**0.99**	0.97	0.99	0.95
B1 + crowd (spkr:10 utt:50)	**0.97**	0.92	0.83	0.93
B2 + crowd (spkr:10 utt:100)	**0.97**	0.94	0.91	0.92
B3 + crowd (spkr:10 utt:200)	**0.93**	0.88	0.92	0.89
C1 + crowd (spkr:15 utt:75)	**0.99**	0.96	0.79	0.96
C2 + crowd (spkr:15 utt:150)	**0.99**	0.95	0.91	0.94
C3 + crowd (spkr:15 utt:300)	**0.96**	0.90	0.90	0.92
A1 + street (spkr:5 utt:25)	**0.86**	0.74	0.69	0.79
A2 + street (spkr:5 utt:50)	**0.78**	0.66	0.69	0.77
A3 + street (spkr:5 utt:100)	**0.86**	0.72	0.84	0.75
B1 + street (spkr:10 utt:50)	**0.84**	0.75	0.62	0.79
B2 + street (spkr:10 utt:100)	**0.75**	0.68	0.66	0.73
B3 + street (spkr:10 utt:200)	**0.72**	0.62	0.71	0.71
C1 + street (spkr:15 utt:75)	**0.77**	0.67	0.60	0.75
C2 + street (spkr:15 utt:150)	**0.68**	0.60	0.61	**0.68**
C3 + street (spkr:15 utt:300)	0.68	0.62	**0.71**	0.68
A1 + party (spkr:5 utt:25)	**0.97**	0.87	0.88	0.95
A2 + party (spkr:5 utt:50)	0.99	0.93	**1.00**	0.87
A3 + party (spkr:5 utt:100)	**1.00**	0.92	0.99	0.96
B1 + party (spkr:10 utt:50)	**0.98**	0.88	0.83	0.95
B2 + party (spkr:10 utt:100)	**0.96**	0.86	0.88	0.95
B3 + party (spkr:10 utt:200)	**0.96**	0.89	0.90	0.92
C1 + party (spkr:15 utt:75)	**0.98**	0.93	0.81	0.94
C2 + party (spkr:15 utt:150)	**0.94**	0.91	0.87	0.92
C3 + party (spkr:15 utt:300)	**0.92**	0.90	0.90	0.90
A1 + station (spkr:5 utt:25)	**0.92**	0.86	0.77	0.87
A2 + station (spkr:5 utt:50)	0.86	0.76	**0.90**	0.85
A3 + station (spkr:5 utt:100)	0.84	0.75	0.86	**0.87**
B1 + station (spkr:10 utt:50)	**0.89**	0.79	0.69	0.86
B2 + station (spkr:10 utt:100)	**0.84**	0.77	0.76	0.86
B3 + station (spkr:10 utt:200)	**0.81**	0.75	0.81	**0.81**
C1 + station (spkr:15 utt:75)	**0.89**	0.79	0.69	0.84
C2 + station (spkr:15 utt:150)	**0.89**	0.74	0.77	0.80
C3 + station (spkr:15 utt:300)	0.81	0.73	**0.83**	**0.83**

2) Comparison with the VB-based method and agglomerative method

The K values determined using the VB-based and agglomerative methods are also listed in Tables 2 and 3. The results obtained by the proposed method were equal or superior to those with the conventional VB-based (**VB**) methods using both the clean and noisy datasets. In particular, the proposed method obtained substantially better performance when the data were very scarce (e.g. A1, B1, C1, T1, and T2). This implies that nested Gibbs sampling-based estimation can adequately estimate the cluster structure from limited data, which is generally difficult to achieve. In fact, the VB-based method cannot model such limited data. To evaluate the effectiveness of a fully Bayesian approach, we also compared the proposed method with the conventional hierarchical agglomerative method (**HAC-GMM**). The proposed method also outperformed the HAC-GMM in most conditions.

3) Computational cost

We now consider the computational cost based on two features: the number of iterations until convergence and the computation required for each epoch. The T-1 dataset (i.e., 24 speakers and 192 utterances; 9.7 min in total) was used for this experiment. The **VB** approach required about 14.8 s on average for one epoch and 12 iterations until it converged (i.e., real-time factor (RTF) of about 0.0031) when an Intel Xeon 3.00 GHz processor was used. However, the proposed nested Gibbs sampling method required about 41.4 s on average for one epoch and about 63 iterations until the maximum logarithmic marginalized likelihood was obtained (i.e., RTF of about 0.0450), whereas the conventional Gibbs sampling method only required about 1.58 s and about 17 iterations until the maximum logarithmic marginalized likelihood was obtained (i.e., RTF of about 0.0005). Figure 6(a) shows the logarithmic marginalized likelihood obtained when the nested Gibbs sampler was applied to dataset A1. We can see that the chain of samples obtained using the nested Gibbs sampler converged within 100 iterations at most. Compared with the conventional Gibbs sampler, the nested Gibbs sampler required more iterations and computations while it obtains substantially better performance. In fact, the computational cost of the nested Gibbs sampler will increase drastically as the number of utterances increases because many iterations are needed during the sampling process. However, the sampling of fLVs can be parallelized, because the posterior distribution of fLVs is calculated independently of the utterances. Thus, we can reduce the computational time by using multi-threading technology.

VI. Conclusion and Future Work

In this study, we proposed a novel method for estimating a mixture-of-mixture model. The proposed nested Gibbs sampler can efficiently avoid local optimum solutions due to its nested sampling procedure, where the structure of its elemental mixture distributions are sampled jointly. We showed that the proposed method can estimate models accurately for speech utterances drawn from complex multi-modal distributions, whereas the results obtained by the conventional Gibbs sampler-based method were

trapped in local optima. The proposed method also outperformed the conventional agglomerative approach in most conditions.

The proposed MoGMMs can build a hierarchical model from multi-level data that comprise frame-wise observations. Some types of real-world data also has the same kind of structure, such as images comprising a set of pixels. In future research, we plan to apply MoGMMs to the image-clustering problem.

Non-parametric Bayesian approaches have recently been attracting the attention as methods for selecting optimal model structures. For example, the nested Dirichlet process mixture model [30] provides a model selection solution for our MoGMMs. In a previous study, we proposed a non-parametric Bayesian version of a mixture-of-mixture model and showed that it was effective in estimating the number of speakers [31, 32]. However, this model was based on the conventional Chinese restaurant process and we employed the conventional Gibbs sampling method, which is readily trapped by the local optima. In future research, we plan to develop a nested Gibbs sampling-based method for such non-parametric Bayesian models.

APPENDIX

In this appendix, we provide detailed descriptions of how to calculate the posterior probabilities for the fLVs and sLVs in equations (31) and (33), which are required for nested Gibbs sampling.

[sLV]

$$
\begin{aligned}
&p(z_u = i|\mathcal{O}, \mathcal{V}_u, \{\mathcal{V}_{\backslash u}, \mathcal{Z}_{\backslash u}\}) \\
&\quad = \frac{p(\mathcal{O}, \mathcal{V}, \mathcal{Z}_{\backslash u}, z_u = i)}{p(\mathcal{O}, \mathcal{V}, \mathcal{Z}_{\backslash u})} \\
&\quad \propto \frac{p(\mathcal{O}, \mathcal{V}, \mathcal{Z}_{\backslash u}, z_u = i)}{p(\mathcal{O}_{\backslash u}, \mathcal{V}_{\backslash u}, \mathcal{Z}_{\backslash u})} \\
&\quad \propto \exp\left\{\log \frac{\Gamma\left(\sum_j \tilde{w}_{i\backslash u,j}\right)}{\Gamma\left(\sum_j \tilde{w}_{i,j}\right)}\right. \\
&\qquad \left. - \beta \sum_j \left(H(\tilde{\boldsymbol{\Psi}}_{i,j}) - H(\tilde{\boldsymbol{\Psi}}_{i\backslash u,j})\right)\right\} \\
&\quad \triangleq \gamma^{\beta}_{z_u = i|\mathcal{V}}.
\end{aligned} \tag{37}
$$

To derive the result using equation (37), we assume that the marginalized likelihood for each complete data $\{\mathbf{o}_{ut}, v_{ut}, z_u\}$ is independent from the others, and use the fact that

$$
\begin{aligned}
p(\mathcal{O}, \mathcal{V}_u, \{\mathcal{V}_{\backslash u}, \mathcal{Z}_{\backslash u}\}) &= p(\mathcal{O}_{\backslash u}, \mathcal{V}_{\backslash u}, \mathcal{Z}_{\backslash u}) \sum_{z_u} p(\mathcal{O}_u, \mathcal{V}_u, z_u) \\
&\propto p(\mathcal{O}_{\backslash u}, \mathcal{V}_{\backslash u}, \mathcal{Z}_{\backslash u}),
\end{aligned} \tag{38}
$$

$H(\tilde{\boldsymbol{\Psi}}_{i,j})$ in equation (37) denotes the logarithmic likelihood of the complete data $\{\mathcal{O}, \mathcal{Z}, \mathcal{V}\}$, which is defined as follows:

$$
\begin{aligned}
H(\tilde{\boldsymbol{\Psi}}_{i,j}) &\triangleq \log p(\mathcal{O}, \mathcal{V}_{u\backslash t} \{\mathcal{V}_{\backslash u}, \mathcal{Z}_{\backslash u}\}, v_{ut} = j, z_u = i) \\
&\propto \log \Gamma(\tilde{w}_{ij}) - \frac{D}{2} \log \tilde{\xi}_{ij} \\
&\quad + D \log \Gamma\left(\frac{\tilde{\eta}_{ij}}{2}\right) - \frac{\tilde{\eta}_{ij}}{2} \sum\nolimits_d \log \tilde{\sigma}_{ij,d},
\end{aligned} \tag{39}
$$

where $\tilde{h}_i$, $\tilde{w}_{ij}$, $\tilde{\xi}_{ij}$, $\tilde{\eta}_{ij}$, $\tilde{\boldsymbol{\mu}}_{ij}$, and $\tilde{\sigma}_{ij,d}$ denote the hyper-parameters of the marginalized likelihood defined in equation (22). We can also obtain the samples of fLVs from these factorized distributions as follows:

[fLVs]

$$
\begin{aligned}
&p(v_{ut} = j|\mathcal{O}, \mathcal{V}_{u\backslash t}, \{\mathcal{V}_{\backslash u}, \mathcal{Z}_{\backslash u}\}, z_u = i) \\
&\quad = \frac{p(\mathcal{O}, \mathcal{V}_{u\backslash t}, \{\mathcal{V}_{\backslash u}, \mathcal{Z}_{\backslash u}\}, v_{ut} = j, z_u = i)}{p(\mathcal{O}, \mathcal{V}_{u\backslash t}, \{\mathcal{V}_{\backslash u}, \mathcal{Z}_{\backslash u}\}, z_u = i)} \\
&\quad \propto \frac{p(\mathcal{O}, \mathcal{V}_{u\backslash t}, \{\mathcal{V}_{\backslash u}, \mathcal{Z}_{\backslash u}\}, v_{ut} = j, z_u = i)}{p(\mathcal{O}_{\backslash\{ut\}}, \mathcal{V}_{\backslash\{ut\}}, \mathcal{Z}_{\backslash u}, z_u = i)} \\
&\quad \propto \exp\left\{-\beta\left(H(\tilde{\boldsymbol{\Psi}}_{i,j}) - H(\tilde{\boldsymbol{\Psi}}_{i,j\backslash t})\right)\right\} \\
&\quad \triangleq \gamma^{\beta}_{v_{ut} = j|z_u = i},
\end{aligned} \tag{40}
$$

where $H(\tilde{\boldsymbol{\Psi}}_{i\backslash u,j})$ $H(\tilde{\boldsymbol{\Psi}}_{i,j\backslash t})$ in equations (37) and (40) denote the logarithmic likelihood of complete data with respect to $\{\mathcal{O}_{\backslash t}, \mathcal{Z}, \mathcal{V}_{\backslash t}\}$ and $\{\mathcal{O}_{\backslash u}, \mathcal{Z}_{\backslash u}, \mathcal{V}_{\backslash u}\}$, respectively.

To derive the result equation (40), we assume that the marginalized likelihood for each complete data $\{\mathbf{o}_{ut}, v_{ut}, z_u\}$ is i.i.d. and we use the fact that

$$
\begin{aligned}
&p(\mathcal{O}, \mathcal{V}_{u\backslash t}, \{\mathcal{V}_{\backslash u}, \mathcal{Z}_{\backslash u}\}, z_u = i) \\
&\quad = p(\mathcal{O}_{\backslash\{ut\}}, \mathcal{V}_{\backslash\{ut\}}, \mathcal{Z}_{\backslash u}, z_u = i) \sum_{v_{ut}} p(\mathbf{o}_{ut}, v_{ut}, z_u = i) \\
&\quad \propto p(\mathcal{O}_{\backslash\{ut\}}, \mathcal{V}_{\backslash\{ut\}}, \mathcal{Z}_{\backslash u}, z_u = i).
\end{aligned} \tag{41}
$$

REFERENCES

[1] Watanabe, S.; Mochihashi, D.; Hori, T.; Nakamura, A.: Gibbs sampling based multi-scale mixture model for speaker clustering, in *ICASSP*, 2011, 4524–4527.

[2] Rabiner, L.; Juang, B.H.: Fundamentals of Speech Recognition. Signal Processing. Prentice-Hall, Upper Saddle River, NJ, 1993.

[3] 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.

[4] Spellman, E.; Vemuri, B.C.; Rao, M.: Using the KL-center for efficient and accurate retrieval of distributions arising from texture images, in *CVPR (1)*, 2005, 111–116.

[5] Bishop, C.M.: Pattern Recognition and Machine Learning, Springer-Verlag, New York, 2006.

[6] McLachlan, G.; Peel, D.: Finite Mixture Models. John Wiley & Sons, New York, 2004.

[7] Andrew, J.L.; McNicholas, P.D.; Sudebi, S.: Model-based classification via mixtures of multivariate t-distributions. Comput. Stat. Data Anal., **55** (1) (2011), 520–529.

[8] Banerjee, A.; Dhillon, I.S.; Ghosh, J.; Sra, S.: Clustering on the unit hypersphere using von Mises–Fisher distributions. J. Mach. Learn. Res., **6** (2005), 1345–1382.

[9] Tang, H.; Chu, S.M.; Huang, T.S.: Generative model-based speaker clustering via mixture of von Mises–Fisher distributions, in *ICASSP*, 2009, 4101–4104.

[10] Marron, J.S.; Wand, M.P.: Exact mean integrated squared error. Ann. Stat., **20** (2) (1992), 712–736.

[11] Lawrence, C.J.; Krzanowski, W.J.: Mixture separation for mixed-mode data. Stat. Comput., **6** (1996), 85–92.

[12] Willse, A.; Boik, R.J.: Identifiable finite mixtures of location models for clustering mixed-mode data. Stat. Comput., **9** (1999), 111–121.

[13] Calo, D.G.; Montanari, A.; Viroli, C.: A hierarchical modeling approach for clustering probability density functions. Comput. Stat. Data Anal., **71** (2014), 79–91.

[14] Vermunt, J.K.: A hierarchical mixture model for clustering three-way data sets. Comput. Stat. Data Anal., **51** (11) (2007), 5368–5376.

[15] Vermunt, J.K.; Magidson, J.: Hierarchical mixture models for nested data structures, in *Classification: The Ubiquitous Challenge*. Springer, Heidelberg, 2005, 240–247.

[16] Dempster, A.P.; Laird, N.M.; Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc.: Ser. B, **39** (1) (1977), 1–38.

[17] Tawara, N.; Ogawa, T.; Watanabe, S.; Kobayashi, T.: Fully Bayesian inference of multi-mixture Gaussian model and its evaluation using speaker clustering, in *ICASSP*, 2012, 5253–5256.

[18] Valente, F.; Motlícek, P.; Vijayasenan, D.: Variational Bayesian speaker diarization of meeting recordings, in *ICASSP*, 2010, 4954–4957.

[19] Valente, F.; Wellekens, C.J.: Variational Bayesian adaptation for speaker clustering, in *ICASSP*, vol. 03, 2005, 965–968.

[20] Teh, Y.W.; Newman, D.; Welling, M.: A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation, in *Advances in Neural Information Processing Systems*, vol. **19**, 2007, 1353–1360.

[21] Blei, D.M.; Ng, A.Y.; Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res., **3** (2003), 993–1022.

[22] Sung, J.; Ghahramani, Z.; Bang, S.: Latent-space variational Bayes. IEEE Trans. PAMI, **30** (12) (2008), 2236–2242.

[23] Teh, Y.W.; Newman, D.; Welling, M.: A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. 2006.

[24] Kirkpatrick, S.; Gelatt, C.D.; Vecchi, M.P.: Optimization by simulated annealing. Science, **220** (4598) (1983), 671–680.

[25] Liu, J.S.: Monte Carlo Strategies in Scientific Computing. Springer, New York, Berlin, Heidelberg, 2008.

[26] Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Fiscus, J.G.; Pallett, D.S.; Dahlgren, N.L.: DARPA TIMIT acoustic phonetic continuous speech corpus CDROM. 1993.

[27] Kawahara, T.; Nanjo, H.; Furui, S.: Automatic transcription of spontaneous lecture speech, in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2001, 186–189.

[28] Shuichi, I.: On recent speech corpora activities in Japan. J. Acoust. Soc. Japan (E), **20** (3) (1999), 163–169.

[29] Solomonoff, A.; Mielke, A.; Schmidt, M.; Gish, H.: Clustering speakers by their voices, in *ICASSP*, 1998, 757–760.

[30] Rodriguez, A.E.G.A.; Dunson, D.B.: The nested Dirichlet process. J. Am. Stat. Assoc., **103** (2008), 1131–1154.

[31] Tawara, N.; Ogawa, T.; Watanabe, S.; Nakamura, A.; Kobayashi, T.: Fully Bayesian speaker clustering based on hierarchically structured utterance-oriented Dirichlet process mixture model, in *INTERSPEECH*, 2012, 5253–5256.

[32] Tawara, N.; Ogawa, T.; Watanabe, S.; Nakamura, A.; Kobayashi, T.: A sampling-based speaker clustering using utterance-oriented Dirichlet process mixture model and its evaluation on large scale data. APSIPA Transactions on Signal and Information Processing, **4** (2015), E16.

Naohiro Tawara received his B.S. and M.S. degrees from Waseda University in Tokyo, Japan in 2010 and 2012. He is currently working as a Research Associate in Waseda University. He is a member of the Acoustical Society of Japan. His research interests include speaker recognition, image processing, and machine learning.

Tetsuji Ogawa received his B.S., M.S., and Ph.D. in electrical engineering from Waseda University in Tokyo, Japan, in 2000, 2002, and 2005. He was a Research Associate from 2004 to 2007 and a Visiting Lecturer in 2007 at Waseda University. He was an Assistant Professor at Waseda Institute for Advanced Study from 2007 to 2012. He has been an Associate Professor at Waseda University and Egypt-Japan University of Science and Technology (E-JUST) since 2012. He was a Visiting Scholar in the Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, from June to September in 2012 and from June to August in 2013, and a Visiting Scholar in Speech Processing Group, Faculty of Information Technology, Brno University of Technology, Czech Republic from June to July in 2014 and May to August in 2015. His research interests include stochastic modeling for pattern recognition, speech enhancement, and speech and speaker recognition. He is a member of the Institute for of Electrical and Electronics Engineering (IEEE), Information Processing Society of Japan (IPSJ) and Acoustic Society of Japan (ASJ). He received the Awaya Prize Young Researcher Award from the ASJ in 2011 and Yamashita SIG Research Award from the IPSJ in 2013.

Shinji Watanabe is a Senior Principal Research Scientist at Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA. He received his Ph.D. from Waseda University, Tokyo, Japan, in 2006. 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 in Georgia institute of technology, Atlanta, GA. His research interests include Bayesian machine learning and speech and spoken language processing. He has been published more than 100 papers in journals and conferences, and received several awards including the Best paper award from the IEICE in 2003. He served an Associate Editor of the IEEE Transactions on Audio Speech and Language Processing, and he is a member of several committees including the IEEE Signal Processing Society Speech and Language Technical Committee and the APSIPA Speech, Language, and Audio Technical Committee.

Tetsunori Kobayashi received B.E., M.E., and Dr.E. degrees from Waseda University, Japan, in 1980, 1982, and 1985, respectively. In 1985, he joined Hosei University where he served as a lecturer and then as an associate professor. In 1991, he moved to Waseda University and has been a professor there since 1997. He was a visiting researcher in MIT's Laboratory for Computer Science, Advanced Telecommunication Laboratory, and NHK's Science and Technical Research Laboratory. His research interests include the basics of speech recognition and synthesis and of image processing and applying them to conversational robots.