MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Blocked Gibbs Sampling Based Multi-Scale Mixture Model for Speaker Clustering on Noisy Data

Tawara, N.; Ogawa, T.; Watanabe, S.; Nakamura, A.; Kobayashi, T.

TR2013-091 September 2013

Abstract

A novel sampling method is proposed for estimating a continuous multi-scale mixture model. The multi-scale mixture models we assume have a hierarchical structure in which each component of the mixture is represented by a Gaussian mixture model (GMM). In speaker modeling from speech, this GMM represents intra-speaker dynamics derived from the difference in the attributes such as phoneme contexts and the existence of non-stationary noise and the mixture of GMMs (MoGMMs) represents inter-speaker dynamics derived from the difference in speakers. Gibbs sampling is a powerful technique to estimate such hierarchically structured models but can easily induce the local optima problem depending on its use especially when the elemental GMMs are complex in structure. To solve this problem, a highly accurate and robust sampling method based on the blocked Gibbs sampling and iterative conditional modes (ICM) is proposed and effectively applied for reducing a singularity solution given in the model with complex multi-modal distributions. In speaker clustering experiments under non-stationary noise, the proposed sampling-based model estimation improved the clustering performance by 17% on average compared to the conventional sampling-based methods.

IEEE International Workshop on Machine Learning for Signal Processing (MLSP)

2013 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 22–25, 2013, SOUTHAPMTON, UK

BLOCKED GIBBS SAMPLING BASED MULTI-SCALE MIXTURE MODEL FOR SPEAKER CLUSTERING ON NOISY DATA

Naohiro Tawara[1], *Tetsuji Ogawa*[1], *Shinji Watanabe*[2], *Atsushi Nakamura*[3], *Tetsunori Kobayashi*[1]

[1] Waseda University, Tokyo, Japan
[2] Mitsubishi Electric Research Laboratories (MERL), MA, USA
[3] NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan

ABSTRACT

A novel sampling method is proposed for estimating a continuous multi-scale mixture model. The multi-scale mixture models we assume have a hierarchical structure in which each component of the mixture is represented by a Gaussian mixture model (GMM). In speaker modeling from speech, this GMM represents intra-speaker dynamics derived from the difference in the attributes such as phoneme contexts and the existence of non-stationary noise and the mixture of GMMs (MoGMMs) represents inter-speaker dynamics derived from the difference in speakers. Gibbs sampling is a powerful technique to estimate such hierarchically structured models but can easily induce the local optima problem depending on its use especially when the elemental GMMs are complex in structure. To solve this problem, a highly accurate and robust sampling method based on the blocked Gibbs sampling and iterative conditional modes (ICM) is proposed and effectively applied for reducing a singularity solution given in the model with complex multi-modal distributions. In speaker clustering experiments under non-stationary noise, the proposed sampling-based model estimation improved the clustering performance by 17% on average compared to the conventional sampling-based methods.

Index Terms— Fully Bayesian approach, blocked Gibbs sampling, iterative conditional modes, multi-scale mixture model, speaker clustering

1. INTRODUCTION

Robust speech modeling is a key for many applications of recognition based on statistical models. Unsupervised speaker modeling from speech, which is also referred to as speaker clustering, plays an important role in speech applications. The difficulty is that speech segments involve multi-scale dynamics, e.g., utterance-level and frame-level dynamics. Utterance-level dynamics, which can be observed in every utterance, are mainly driven by the acoustic changes in global attributes such as speakers, their emotions, and spoken topics. Frame-level dynamics, on the other hand, which can be observed in short-time analysis in a period of a few dozen microseconds, are mainly derived from local acoustic variations such as phoneme contexts and background non-stationary noise.

Latent variable models are effective in modeling such hierarchically structured data. In the research field of natural language processing, for example, probabilistic latent semantic analysis (pLSA) [1] and its fully Bayesian extension, latent Dirichlet allocation (LDA) [2], are successful approaches to expressing multiple scales on discrete data. For speech data as well, which are ordinarily composed of multivariate continuous data, multi-scale mixture models have been used to represent the hierarchical structure of speech [3, 4, 5].

Bayesian estimation can play an essential role in robust model estimation on real data. Many researchers have attempted to use Bayesian approaches for speech modeling; the maximum a posterior (MAP) based method [6] and variational Bayesian (VB) based method [7] were applied to speaker recognition [8] and speaker clustering [3, 9]. These approaches are based on deterministic algorithms using the expectation maximization (EM) algorithm.

In contrast, we have focused on the stochastic approach based on the Markov chain Monte Carlo (MCMC) approach [4, 5]. We applied a simple collapsed Gibbs sampler to obtain the utterance- and frame-level latent variables from their joint posterior distribution. This approach involves first sampling the frame-level latent variables (fLVs) and then sampling the utterance-level latent variables (uLVs). Both MCMC- and VB-based approaches are based on the estimation of posterior distribution of latent variables. The MCMC-based approach, however, has an obvious advantage that the posterior distribution can be marginalized over the model parameters and the instantiated values of the latent variables can be directly obtained. The VB-based approach, on the other hand, needs to evaluate the posterior distribution of model parameters and it can provide the severe over-fitting problem in the case where the number of spoken utterances is small [5]. However, this sampler has a severe restriction in that the sampling step of uLVs is strictly determined by the values of fLVs that are obtained in the previous sampling step. This restriction can induce the local optima problem in uLVs because the uLVs estimated in every iterations can be highly correlated. To relax this constraint, we propose a novel sampling method based on blocked Gibbs sampling, which samples both uLVs and fLVs at the same time. This sampler makes it possible to efficiently evaluate the enormous combination of fLVs and uLVs and therefore find a more appropriate solution than the conventional sampling method. We evaluated the effectiveness of the proposed sampling method using speaker clustering experiments under non-stationary noise.

2. FORMULATION

We explain a multi-scale mixture model with GMMs. 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 utterance, $\mathbf{O}_u \triangleq \{\mathbf{o}_{ut}\}_{t=1}^{T_u}$ be the u-th utterance that comprises the T_u observation vectors, and $\mathcal{O} \triangleq \{\mathbf{O}_u\}_{u=1}^{U}$ be a set of U utterances.

We introduce a generative model in which all utterances $\mathcal{O}$ are generated from a mixture of GMMs (MoGMMs) in which each GMM represents the speaker characteristics (i.e., intra-speaker vari-

978-1-4799-1180-6/13/$31.00

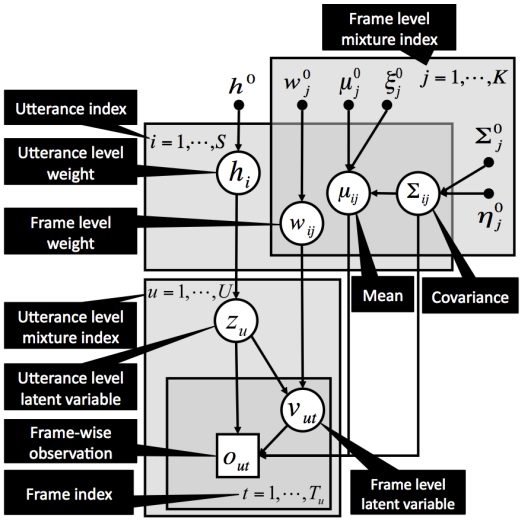

Fig. 1. *Graphical representations of multi-scale mixture model.*

ability) and a mixture of these GMMs represents the entire speaker space (i.e., inter-speaker variability). We call this multi-scale mixture model.

To deal with this hierarchical mixture model, we introduce two types of latent variables; $\mathcal{Z} = \{z_u\}_{u=1}^{U}$ represents the utterance-level latent variables, each of which identifies the MoGMMs component (i.e., speaker distribution) to which the u-th utterance is assigned; and $\mathcal{V} = \{\{v_{ut}\}_{t=1}^{T_u}\}_{u=1}^{U}$ represents the frame-level latent variables, each of which identifies the intra-speaker GMM component to which the t-th frame-wise observation in the u-th utterance is assigned. By introducing these latent variables, we can describe the conditional probability of all utterances given the latent variables as follows [1]:

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

$$P(\mathcal{V}|\mathcal{Z},\mathbf{w}) = \prod_{u=1}^{U}\prod_{t=1}^{T_u}\prod_{i=1}^{S}\prod_{j=1}^{K} w_{ij}^{\delta(v_{ut},j)\delta(z_u,i)}, \quad (2)$$

$$P(\mathcal{Z}|\mathbf{h}) = \prod_{u=1}^{U}\prod_{i=1}^{S} h_i^{\delta(z_u,i)}, \quad (3)$$

where $\boldsymbol{\Theta} \triangleq \{\{h_i\},\{w_{ij}\},\{\boldsymbol{\mu}_{ij}\},\{\boldsymbol{\Sigma}_{ij}\}\}$ denote the weight, mean vector, and covariance matrix of the component of the intra-speaker GMM, respectively; $\delta(a,b)$, Kronecker's delta, which takes one if $a = b$ and zero otherwise. Note that we have assumed $\boldsymbol{\Sigma}_{ij}$ is a diagonal covariance matrix whose (d,d)-th element is represented by $\sigma_{ij,d}$. In a Bayesian approach, the conjugate prior distributions of the parameters are often introduced as follows:

$$p(\boldsymbol{\Theta}|\boldsymbol{\Theta}^0) = \begin{cases} \mathbf{h} & \sim \mathcal{D}(\mathbf{h}^0) \\ \mathbf{w}_i & \sim \mathcal{D}(\mathbf{w}^0) \\ \{\mu_{ij,d},\sigma_{ij,d}\} & \sim \mathcal{NG}(\xi^0,\eta^0,\mu_{j,d}^0,\sigma_{j,d}^0) \end{cases} \quad (4)$$

where $\mathcal{D}\left(\mathbf{h}^0\right)$ and $\mathcal{D}\left(\mathbf{w}^0\right)$ denote the Dirichlet distribution with a hyperparameter $\mathbf{h}^0$ and $\mathbf{w}^0$, respectively. $\mathcal{NG}\left(\xi^0,\eta^0,\mu_{j,d}^0,\sigma_{j,d}^0\right)$ denotes the Gaussian-Gamma distribution with hyperparameters ξ^0, η^0, $\mu_{j,d}^0$, and $\sigma_{j,d}^0$. Figure 1 shows a graphical representation of this model.

[1] We use notation $p(\cdot)$ for the continuous probability function and $P(\cdot)$ for the discrete probability function.

3. MODEL INFERENCE

The key issue in Bayesian learning in the latent variable model is estimating the posterior distribution of latent variables, $p(\mathcal{V},\mathcal{Z},\boldsymbol{\Theta}|\mathcal{O})$. When the multi-scale mixture model described in the previous section is used, the speaker clustering problem is reduced to the problem of finding the optimal assignments of the utterances to the speakers (i.e., the utterance-level latent variables $\mathcal{Z}$) so as to maximize the posterior probability. However, the direct optimization of posterior distribution $p(\mathcal{V},\mathcal{Z},\boldsymbol{\Theta}|\mathcal{O})$ is infeasible, and therefore, some approximation is required.

In [3], the VB approach is introduced in order to find the optimal posterior distribution. The VB-based methods, however, can suffer from overfitting of the posterior hyperparameters for cases with a limited number of observations $\mathcal{O}$ is limited [5]. To address this problem, we can marginalize out the model parameters, $\boldsymbol{\Theta}$, from the joint distribution and obtain: $P(\mathcal{V},\mathcal{Z}|\mathcal{O}) \propto \int p(\mathcal{O},\mathcal{V},\mathcal{Z},\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}$. In this case, since estimation of the model parameters is no longer needed, we can robustly estimate the posterior distribution even for a limited amount of data. In contrast, with the VB method, it is usually infeasible to marginalize them without any approximation [10].

MCMC has been introduced as an alternative to VB approximation [4, 5]. In this approach, we can directly obtain the latent variables from their posterior distribution ($P(\mathcal{V},\mathcal{Z}|\mathcal{O})$).

In the rest of this section, we review the conventional model estimation method using Gibbs sampling [4, 5] (**3.1**). Then, we discuss the problem induced by the conventional method and propose a more accurate sampling method using blocked Gibbs sampling (**3.2** and **3.3**).

3.1. Collapsed Gibbs sampling approach

The marginalized likelihood for the complete data $p(\mathcal{O},\mathcal{V},\mathcal{Z})$ can be analytically obtained by substituting Eqs. (1) to (3) for Eq. (4) into the following integral equation:

$$\begin{aligned} & p(\mathcal{O},\mathcal{V},\mathcal{Z}) \\ &= \int p(\mathcal{O},\mathcal{V},\mathcal{Z}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta} \\ &= \frac{\Gamma(h^0)\prod_i \Gamma(\tilde{h}_i)}{\Gamma(h^0/S)^S\Gamma(\sum_i \tilde{h}_i)} \cdot \prod_i \frac{\Gamma(\sum_j w_j^0)\prod_j \Gamma(\tilde{w}_{ij})}{\prod_j \Gamma(w_j^0)\Gamma(\sum_j \tilde{w}_{ij})} \\ &\quad \cdot \prod_{i,j}(2\pi)^{-\frac{n_{ij}D}{2}} \frac{(\xi^0)^{\frac{D}{2}}\left(\Gamma\left(\frac{\eta_j^0}{2}\right)\right)^{-D}\left(\prod_d \sigma_{j,dd}^0\right)^{\frac{\eta_j^0}{2}}}{(\tilde{\xi}_{ij})^{\frac{D}{2}}\left(\Gamma\left(\frac{\tilde{\eta}_{ij}}{2}\right)\right)^{-D}\left(\prod_d \tilde{\sigma}_{ij,dd}\right)^{\frac{\tilde{\eta}_{ij}}{2}}}. \end{aligned} \quad (5)$$

Here, $\tilde{\boldsymbol{\Theta}}_{ij} \triangleq \left\{\tilde{h}_i,\tilde{w}_{ij},\tilde{\xi}_{ij},\tilde{\eta}_{ij},\tilde{\mu}_{ij,d},\tilde{\sigma}_{ij,d}\right\}$ denote the hyperparameters of the joint distribution as follows:

$$\begin{cases} \tilde{h}_i & = h^0 + c_i, \\ \tilde{w}_{ij} & = w_j^0 + n_{ij}, \\ \tilde{\xi}_{ij} & = \xi^0 + n_{ij}, \\ \tilde{\eta}_{ij} & = \eta^0 + n_{ij}, \\ \tilde{\boldsymbol{\mu}}_{ij} & = \frac{\xi^0\boldsymbol{\mu}_j^0 + \mathbf{m}_{ij}}{\tilde{\xi}_{ij}}, \\ \tilde{\sigma}_{ij,d} & = \sigma_{j,d}^0 + r_{ij,d} + \xi^0(\mu_{j,d}^0)^2 - \tilde{\xi}_{ij}(\tilde{\mu}_{ij,d})^2, \end{cases} \quad (6)$$

Algorithm 1 Conventional Gibbs sampling based model estimation. $\mathcal{M}(\cdot)$ denotes the multinomial distribution.

1: Initialize $\{z_u, v_{ut} : u = 1, \cdots, U, t = 1, \cdots, T_u\}$.
2: **repeat**
3: **for all** *utterances* u and *frames* t **do**
4: **for all** *components* j **do**
5: Compute $\gamma_{v_{ut}=j|z_u=i,\mathcal{V}_t,\mathcal{Z}_{\setminus u}}$ by Eq. (8).
6: **end for**
7: Draw frame-level latent variable (fLV), v^*_{ut}, from its conditional posterior distribution, $\mathcal{M}\left(\frac{\gamma_{v_{ut}=j|z_u=i,\mathcal{V}_t,\mathcal{Z}_{\setminus u}}}{\sum_j \gamma_{v_{ut}=j|z_u=i,\mathcal{V}_t,\mathcal{Z}_{\setminus u}}}\right)$.
8: **end for**
9: **for all** *utterances* u **do**
10: **for all** *speakers* i **do**
11: Compute $\gamma_{z_u=i|\mathcal{V},\mathcal{Z}_{\setminus u}}$ by Eq. (9)
12: **end for**
13: Draw utterance-level latent variable (uLV), z^*_u, from its conditional posterior distribution, $\mathcal{M}\left(\frac{\gamma_{z_u=\cdot|\mathcal{V},\mathcal{Z}_{\setminus u}}}{\sum_i \gamma_{z_u=i|\mathcal{V},\mathcal{Z}_{\setminus u}}}\right)$.
14: **end for**
15: **until** some condition is met

where c_i, n_{ij}, $\mathbf{m}_{ij}$, and $r_{ij.d}$ denote the zero-th, first, and second order sufficient statistics as follows:

$$\begin{cases} c_i & = \sum_u \delta(z_u, i), \\ n_{ij} & = \sum_{u,t} \delta(z_u, i) \cdot \delta(v_{ut}, j), \\ \mathbf{m}_{ij} & = \sum_{u,t} \delta(z_u, i) \cdot \delta(v_{ut}, j) \cdot \mathbf{o}_{ut}, \\ r_{ij,d} & = \sum_{u,t} \delta(z_u, i) \cdot \delta(v_{ut}, j) \cdot (o_{ut,d})^2, \end{cases} \tag{7}$$

where c_i denotes the number of utterances assigned to the i-th component of the entire speaker MoGMMs; n_{ij}, the number of frames assigned to the j-th component of the i-th intra-speaker GMM of the MoGMMs; and $\mathbf{m}_{ij}$ and r_{ij}, the first and second sufficient statistics, respectively.

The MCMC-based method estimates the utterance- and frame-level latent variables by directly obtaining the samples of the latent variables from their posterior distribution. In [4, 5], we applied collapsed [1] Gibbs sampling [11]. In each step of collapsed Gibbs sampling, we replace the value of the latent variable (e.g., z_u) with a value drawn from the distribution $p(z_u|\mathcal{Z}_{\setminus u})$, where $\mathcal{Z}_{\setminus u}$ denotes the set of variables but with z_u omitted (i.e. $\mathcal{Z}_{\setminus u} = \{z_{u'}|u' \neq u\}$). In the case of a multi-scale mixture model, the frame- and utterance-level latent variables are respectively sampled from the conditional posterior distribution as follows:

[Frame-level latent variables]

$$\begin{aligned} &\gamma_{v_{ut}=j'|z_u=i,\mathcal{V}_t,\mathcal{Z}_{\setminus u}} \\ &= p(v_{ut} = j'|\mathcal{O}, \mathcal{V}_{\setminus t}, \mathcal{Z}_{\setminus u}, z_u = i) \\ &= \frac{p(\mathcal{O}, \mathcal{V}_{\setminus t}, v_{ut} = j', \mathcal{Z}_{\setminus u}, z_u = i)}{p(\mathcal{O}_{\setminus t}, \mathcal{V}_{\setminus t}, \mathcal{Z}_{\setminus u}, z_u = i)} \\ &= \exp\left(g_{ij'}(\tilde{\mathbf{\Psi}}_{i,j'}) - g_{ij'}(\tilde{\mathbf{\Psi}}_{i,j'\setminus t})\right) \end{aligned} \tag{8}$$

[Utterance-level latent variable]

$$\begin{aligned} &\gamma_{z_u=i'|\mathcal{V},\mathcal{Z}_{\setminus u}} \\ &= p(z_u = i'|\mathcal{O}, \mathcal{V}, \mathcal{Z}_{\setminus u}) \\ &= \frac{p(\mathcal{O}, \mathcal{V}, \mathcal{Z}_{\setminus u}, z_u = i')}{p(\mathcal{O}_{\setminus u}, \mathcal{V}_{\setminus u}, \mathcal{Z}_{\setminus u})} \\ &= \exp\left(\log \frac{\Gamma(\sum_j \tilde{w}_{i'\setminus u,j})}{\Gamma(\sum_j \tilde{w}_{i',j})} + \sum\nolimits_j \left(g_{ij}(\tilde{\mathbf{\Psi}}_{i',j}) - g_{ij}(\tilde{\mathbf{\Psi}}_{i'\setminus u,j})\right)\right) \end{aligned} \tag{9}$$

where $g_{ij}(\tilde{\mathbf{\Psi}}_{i,j})$ denotes the joint logarithmic probability described as:

$$\begin{aligned} g_{ij}(\tilde{\mathbf{\Psi}}_{i,j}) &\triangleq \log p(\mathcal{O}, \mathcal{V}_{\setminus ut}, v_{ut} = j, \mathcal{Z}_{\setminus u}, z_u = i) \\ &\propto \log \Gamma(\tilde{w}_{ij}) - \frac{D}{2} \log \tilde{\xi}_{ij} \\ &\quad + D \log \Gamma\left(\frac{\tilde{\eta}_{ij}}{2}\right) - \frac{\tilde{\eta}_{ij}}{2} \sum\nolimits_d \log \tilde{\sigma}_{ij,d}, \end{aligned} \tag{10}$$

where $\tilde{h}_i$, $\tilde{w}_{ij}$, $\tilde{\xi}_{ij}$, $\tilde{\eta}_{ij}$, $\tilde{\boldsymbol{\mu}}_{ij}$, and $\tilde{\sigma}_{ij,d}$ are described in Eq. (6), and D denotes the dimensionality of observation vectors. Here, $g_{ij}(\tilde{\mathbf{\Psi}}_{i,j\setminus t})$ is computed using $\mathcal{O}_{\setminus t}$, $\mathcal{Z}$, and $\mathcal{V}_{\setminus t}$.

This sampling process is iterated across all frame- and utterance-level latent variables. For the multi-scale mixture model used, the collapsed Gibbs sampling procedure is carried out as Algorithm 1. $\mathcal{M}(\cdot)$ denotes the multinomial distribution.

3.2. Blocked Gibbs sampling

As described in 3.1, the conventional method using collapsed Gibbs sampling carries out sampling of frame-level latent variables (fLVs) followed by sampling of utterance-level latent variables (uLVs). In this case, the values of uLVs are drawn from their posterior distribution conditioned by the sampled values of fLVs. This indicates that the posterior distribution of uLV described in Eq. (9) is evaluated using the *identical* values of fLVs (i.e., assignments of every frames to the components of GMM) for all intra-speaker GMMs clusters. The restriction in which the sampled uLVs are evaluated using the fixed fLVs can result in a serious problem in that the convergence speed might be slow, especially for complex distributions that are composed of a large number of mixtures. Since a substantial quantity of samples is needed in order to evaluate a lot of combinations of the fLVs and uLVs, the sampling process necessarily needs substantial iterations until the chains of samples are converged to the true posterior distribution.

To avoid this problem, we propose a new sampling method that can evaluate a larger number of hypotheses at each iteration. In each step of the proposed sampling method, once the u-th utterance is chosen, the set of fLVs and uLV, $\{z_u, \mathcal{V}_u\}$, is drawn from their joint and posterior distribution, $P(z_u, \mathcal{V}_u|\mathcal{Z}_{\setminus u}, \mathcal{V}_{\setminus u}, \mathcal{O})$. Here, this distribution can be factorized as

$$\begin{aligned} &P(z_u, \mathcal{V}_u|\mathcal{Z}_{\setminus u}, \mathcal{V}_{\setminus u}, \mathcal{O}) \\ &= P(z_u|\mathcal{Z}_{\setminus u}, \mathcal{V}_u, \mathcal{V}_{\setminus u}, \mathcal{O}) P(\mathcal{V}_u|z_u, \mathcal{Z}_{\setminus u}, \mathcal{V}_{\setminus u}, \mathcal{O}). \end{aligned} \tag{11}$$

We can obtain the samples from this factorized distribution by ancestral sampling [12]. In this case, we attempt to introduce another Gibbs sampler that draws the values of fLVs from the second term on the right side of Eq. (11) as

$$\mathcal{V}^*_u \sim P(\mathcal{V}^*_u|z_u, \mathcal{Z}_{\setminus u}, \mathcal{V}_{\setminus u}, \mathcal{O}). \tag{12}$$

[1] The term *collapsed* means that samples are drawn from the marginalized distribution with respect to model parameter $\mathbf{\Theta}$.

Algorithm 2 Blocked Gibbs sampling-based model estimation with iterated conditional modes (ICM) approximation. $\mathcal{M}(\cdot)$ denotes multinomial distribution.

1: Initialize $\{z_u, v_{ut} : u = 1, \cdots, U, t = 1, \cdots, T_u\}$.
2: **repeat**
3: **for all** *utterances* u **do**
4: **for all** *speakers* i **do**
5: **for all** *frames* t **do**
6: **for all** *components* j **do**
7: Compute $\gamma_{v_{ut}=j|z_u=i,\mathcal{V}_t,\mathcal{Z}_{\setminus u}}$ by Eq. (8).
8: **end for**
9: Decide the value of fLV, v^*_{ut}, from its posterior probability by $v^*_{ut} = \arg\max_j \gamma_{v_{ut}=j}|z^*_u$
10: **end for**
11: Compute $\gamma_{z_u=i|\mathcal{V}^*,\mathcal{Z}_{\setminus u}}$ by Eq. (9) conditioning on the sampled fLVs, $\{v^*_{ut}\}_{t=1}^{T_u}$.
12: **end for**
13: **end for**
14: **for all** *utterances* u **do**
15: Draw the value of uLV, z^*_u, from its posterior distribution by $z^*_u \sim \mathcal{M}\left(\frac{\gamma_{z_u=\cdot|\mathcal{V}^*,\mathcal{Z}_{\setminus u}}}{\sum_i \gamma_{z_u=i|\mathcal{V}^*,\mathcal{Z}_{\setminus u}}}\right)$
16: **end for**
17: **until** some condition is met

This Gibbs sampler can be constructed using the posterior distribution of fLVs of the u-th utterance given fLVs of the remaining utterances. Then, the posterior distribution of uLV conditioned by the sampled value of fLVs, $\mathcal{V}^*_u$, is evaluated and a uLV is drawn from its posterior distribution as

$$z_u \sim P(z_u|\mathcal{Z}_{\setminus u}, \mathcal{V}^*_u, \mathcal{V}_{\setminus u}, \mathcal{O}). \tag{13}$$

An essential difference between the proposed and the conventional sampler is that the proposed sampler simultaneously draws the fLVs and uLV by nesting the Gibbs sampler for uLV and that for fLV, whereas the conventional sampler alternately draws them. It should be noted that the proposed sampler is regarded as a kind of blocked Gibbs sampling [13] because it samples a group of variables.

3.3. Greedy approximation for sampling of frame-level latent variables in blocked Gibbs sampler

Trace plots of the K values, which are frequently applied as an evaluation measure of speaker clustering, are shown in Figure 2 (a). Also shown are marginalized likelihoods computed using the samples obtained from the aforementioned blocked Gibbs sampler. In can be seen in this figure that the K values and marginalized likelihoods decrease at around the fourth iteration. We discuss the reason for this by showing the mean vectors of 26-dimensional acoustic feature parameters that are assigned to the components of GMM representing a speaker in Figure 3. This figure describes the mean vectors obtained at the (1) fourth and (2) fifth iterations. In these examples, an intra-speaker GMM has four components. This figure indicates that all components in this GMM have degenerated to the identical distribution. This singular solution is undesirable because the data assigned to the corresponding single speaker cluster should essentially be drawn from the multi-modal distribution. This should be attributed to the properties of sampling techniques e.g., the stochastic decision process randomly assigns a frame-level feature vector to a component when the probabilities of the frame-level data being assigned to all components are almost the same. This problem would often occur in the sampling of fLVs because each speaker has relatively small variability on the feature space and the multiple components are unduly determined as the identical component even though those components should be distinguished. We can avoid this problem by taking an enormous number of chains in Gibbs sampling with respect to fLVs. However, this is usually infeasible from the viewpoint of computational costs. In contrast, we attempt to address this problem by applying the iterated conditional modes (ICM) algorithm [14] rather than using strict Gibbs sampling to estimate fLVs described in Eq. (12). ICM is regarded as a greedy approximation of Gibbs sampling. When using ICM, a point estimate of fLV is given by the maximum of the conditional distribution instead of drawing a sample from the corresponding conditional distribution at each stage of sampling the fLVs. Namely, we determine the value of fLV in the u-th utterance by

$$v^*_{ut} = \arg\max_j \gamma_{v_{ut}=j}|z^*_u. \tag{14}$$

This ICM-based approximation can deterministically assign frame-level feature vectors to an adequate component even if the posterior probabilities of the frame-level features being assigned to components are almost the same.

Algorithm 2 is the algorithm of the proposed model estimation using the blocked Gibbs sampler with the ICM approximation. It should be noted that an ICM algorithm is substituted for the Gibbs sampling procedure for obtaining the samples of fLVs in the line 9.

Figure 2 (b) shows the trace plots of the K values and marginalized likelihoods obtained from blocked Gibbs sampling with ICM approximation. We can see from this figure that the singular solution can be avoided by introducing the ICM to the sampling of fLVs.

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

4. EXPERIMENTS

We compared three model estimation methods as follows:

- **b-Gibbs (proposed):** MCMC-based model estimation with the proposed blocked Gibbs sampler
- **Gibbs:** MCMC-based model estimation with the conventional Gibbs sampler [4, 5]
- **VB:** VB-based model estimation [3]

Experimental comparisons in terms of speaker clustering accuracy were carried out using the TIMIT acoustic-phonetic continuous speech corpus (TIMIT) and the corpus of spontaneous Japanese (CSJ).

4.1. Evaluation conditions

All experiments were conducted using eight evaluation sets obtained from TIMIT and CSJ. Table 1 lists the number of speakers and ut-

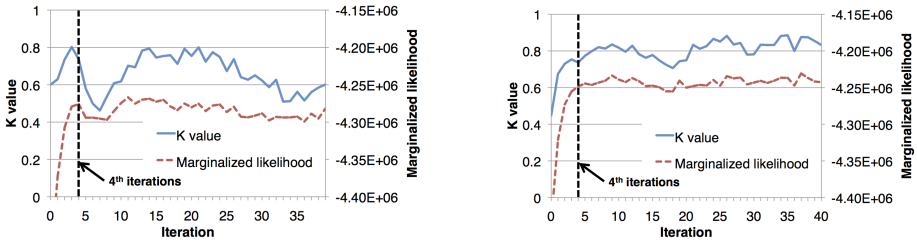

Fig. 2. *K values and logarithmic marginalized likelihood obtained by (a) blocked Gibbs sampling and (b) iterated conditional modes (ICM).*

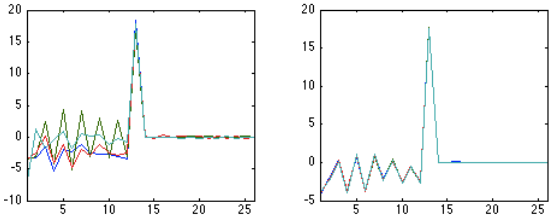

Fig. 3. *Mean vectors of 26-dimensional acoustic feature parameters (MFCCs) assigned to four Gaussian components of a specific speaker GMM at (1) fourth and (2) fifth iterations. Each line corresponds to an assignment to each Gaussian component.*

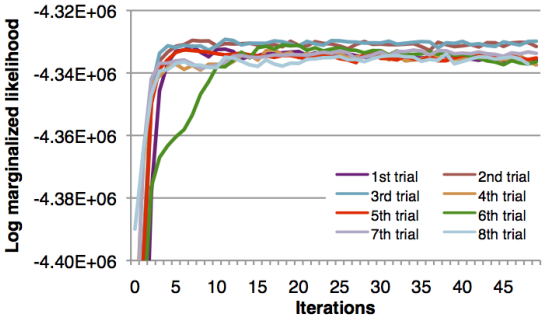

Fig. 4. *Logarithmic marginalized likelihood obtained from blocked Gibbs w/ ICM on T1. Eight lines correspond to the results of eight trials with different seeds.*

terances in those evaluation sets used. T1 and T2 were constructed using TIMIT. T1 corresponds to the core test set of TIMIT, which includes 192 utterances spoken by 24 speakers. T2 was the complete test set, which includes 1,152 utterances spoken by 144 speakers. In this case, there are no overlaps between T1 and T2. The remaining six evaluation sets were constructed using CSJ as follows: all of the lecture speech in CSJ were divided into utterance units on the basis of silence segments in their transcriptions; then, 5 speakers were randomly selected; and their 5, 10, and 20 utterances were chosen for A1, A2 and A3. In the same way, we randomly selected 10 different speakers and their 5, 10, and 20 utterances for B1, B2 and B3. We evaluated five combinations of different speakers on each data set. The resulting performance for each data set is the average for these five combinations. The speech data from TIMIT and CSJ are basically uncorrupted by noise. In addition to those clean speech data, we used noisy speech data that were developed by overlapping each utterance with two types of non-stationary noise at a signal-to-noise ratio (SNR) of about 10 dB. The speech data were sampled at 16 kHz and quantized into 16-bit data. We used 26-dimensional acoustic feature parameters that consisted of 12-dimensional mel-frequency cepstral coefficients (MFCCs) with log energy and their Δ parameters. The frame length and frame shift were 25 ms and 10 ms, respectively.

The evaluation criterion we applied in speaker clustering was the K value, which is the geometric mean of the average speaker purity and average cluster purity [15].

We conducted the same experiments but with different seeds eight times. Then, we evaluated the marginalized likelihood for each result and selected the result with the highest likelihood.

The hyperparameters in Eq. (6) were set as follows: $w^0 = 1$ and $\mathbf{w}^{(0)} = \{w^0, \cdots, w^0\}$ for all components; $h^0 = 1$ and $\mathbf{h}^{(0)} = \{h^0, \cdots, h^0\}$ for all clusters; $\eta^{(0)} = 1$ and $\xi^{(0)} = 1$; $\boldsymbol{\mu}^{(0)}$ and $\boldsymbol{\Sigma}^{(0)}$ were set to the mean vectors and covariance matrices estimated from the whole dataset, respectively. The number of mixtures in the intra-speaker GMMs was set to four for T1 and T2, and eight for the remaining datasets. We randomly initialized both the utterance- and frame-level latent variables.

4.2. Results

Table 2 lists the K values given by three estimation methods for clean speech data. Here, note that the data from CSJ (i.e., A1 to B3) follow an approximately unimodal distribution, and those from TIMIT (i.e., T1 and T2) follow a multi-modal distribution. The conventional VB-based method (**VB**) gave relatively worse performance all the datasets. This implies that the number of data is small for **VB** in all of these datasets. On the other hand, the proposed sampler-based estimation (**b-Gibbs**) and the conventional sampler-based estimation (**Gibbs**) adequately model the all data from CSJ (i.e., A1 to B3), whereas the conventional **Gibbs** failed to model the data from TIMIT (i.e., T1 and especially T2). Table 3 lists the results for noisy data. We can assume that those noisy data follow a multi-

Table 2. K value for clean test sets.

Evaluation data	b-Gibbs	Gibbs	VB
T1 (spkr:24 utt:192)	**0.87**	0.81	0.71
T2 (spkr:144 utt:1152)	**0.74**	0.52	0.41
A1 (spkr:5 utt:25)	**0.99**	0.92	0.88
A2 (spkr:5 utt:50)	**0.99**	0.91	0.95
A3 (spkr:5 utt:100)	**1.00**	0.90	0.98
B1 (spkr:10 utt:50)	0.88	**0.89**	0.73
B2 (spkr:10 utt:100)	**0.95**	0.90	0.76
B3 (spkr:10 utt:200)	**0.97**	0.90	0.80

modal distribution because the noise overlapped are non-stationary. This table shows that the proposed **b-Gibbs** outperformed the conventional **Gibbs** and **VB**, irrespective of the evaluation sets also under the noisy conditions. These results imply that the conventional **Gibbs**, cannot model the data drawn from complex multi-modal distributions (T1 and T2 for clean data and A1 to B3 for noisy data) accurately while it can model the data drawn from simple unimodal distributions (A1 to B3 for clean data). In contrast, the proposed **b-Gibbs** can adequately model the data drawn from both of them.

Next, we discuss the computational cost. In the experiment for use in the T1 dataset (i.e., 24 speakers and 192 utterances), the **VB** took about 14.8 seconds on average in one iteration when using an Intel Xeon 3.00GHz processor. The proposed **b-Gibbs**, on the other hand, took about 41.4 seconds on average in one iteration whereas the conventional **Gibbs** took about 1.58 seconds. It should be noted that both **Gibbs** and **b-Gibbs** required few iterations for convergence, although MCMC-based methods generally require a lot of iterations compared with the VB-based method. Figure 4 shows the logarithmic marginalized likelihood obtained from b-Gibbs on T1. We can see from this figure that the likelihoods are converged after twenty iterations at most. This fast convergence comes from the effect of collapsing the model parameters.

As aforementioned, the proposed **b-Gibbs** requires heavier computation in each iterations, although it achieves higher accuracy than the conventional Gibbs and VB approaches. In addition, the computational cost of **b-Gibbs** will drastically increase as the number of utterances increases because a lot of iterations are needed in the sampling process. Fortunately, it is easy to parallelize the sampling of fLVs because the calculation of the posterior distribution of fLVs is independent with respect to the utterances. We can therefore reduce the computational time by using the parallelized techniques such as general purpose graphical processing unit (GPGPU) and multi-threading technologies.

5. CONCLUSION

We proposed a method of estimating a multi-scale mixture model using blocked Gibbs sampling and ICM. We showed that the proposed method could estimate the model accurately for the speech utterances drawn from a complex multi-modal distribution while the results from the conventional Gibbs sampler-based method got trapped in a local optima.

We have studied a non-parametric Bayesian version of multi-scale mixture model and showed that that model was effective in estimating the number of speakers [16]. This model, however, is based on the conventional Gibbs sampling. We would therefore like to introduce a blocked Gibbs sampling based method for this non-parametric Bayesian model.

Table 3. K value for noisy test sets.

Evaluation data	b-Gibbs	Gibbs	VB
A1 +noise1 (spkr:5 utt:25)	**0.89**	0.67	0.64
A2 +noise1 (spkr:5 utt:50)	**0.88**	0.71	0.72
A3 +noise1 (spkr:5 utt:100)	**0.84**	0.67	0.74
B1 +noise1 (spkr:10 utt:50)	**0.75**	0.65	0.57
B2 +noise1 (spkr:10 utt:100)	**0.75**	0.66	0.62
B3 +noise1 (spkr:10 utt:200)	**0.77**	0.69	0.74
A1 +noise2 (spkr:5 utt:25)	**0.84**	0.71	0.53
A2 +noise2 (spkr:5 utt:50)	**0.80**	0.66	0.63
A3 +noise2 (spkr:5 utt:100)	**0.88**	0.68	0.72
B1 +noise2 (spkr:10 utt:50)	**0.77**	0.72	0.56
B2 +noise2 (spkr:10 utt:100)	**0.75**	0.61	0.63
B3 +noise2 (spkr:10 utt:200)	**0.74**	0.63	0.71

6. REFERENCES

[1] Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.

[2] David M. Blei et al. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[3] Fabio Valente and Christian J Wellekens. Variational bayesian adaptation for speaker clustering. In *ICASSP*, pages 965–968, 2005.

[4] Shinji Watanabe et al. Gibbs sampling based multi-scale mixture model for speaker clustering. In *ICASSP*, pages 4524–4527, 2011.

[5] Naohiro Tawara et al. Fully bayesian inference of multi-mixture gaussian model and its evaluation using speaker clustering. In *ICASSP*, pages 5253–5256, 2012.

[6] Jean luc Gauvain and Chin hui Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Trans. Speech Audio Process.*, 2:291–298, 1994.

[7] Shinji Watanabe et al. Variational bayesian estimation and clustering for speech recognition. *IEEE Trans. Speech Audio Process.*, 12:365–381, 2004.

[8] Douglas A. Reynolds et al. Speaker verification using adapted gaussian mixture models. In *Digital Signal Processing*, page 2000, 2000.

[9] Fabio Valente et al. Variational bayesian speaker diarization of meeting recordings. In *ICASSP*, pages 4954–4957, 2010.

[10] Jaemo Sung et al. Latent-space variational bayes. *IEEE Trans. on PAMI*, 30(12):2236–2242, 2008.

[11] Jun S. Liu. *Monte Carlo strategies in scientific computing*. Springer, corrected edition, January 2008.

[12] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.

[13] Claus Skaanning Jensen and Augustine Kong. Blocking gibbs sampling in very large probabilistic expert systems. *Internat. J. Human-Computer Studies*, 42:647–666, 1995.

[14] Josef Kittler and J. Föglein. Contextual classification of multispectral pixel data. *Image Vision Comput.*, 2(1):13–29, 1984.

[15] Alex Solomonoff et al. Clustering speakers by their voices. In *ICASSP*, pages 757–760, 1998.

[16] Naohiro Tawara et al. Fully bayesian speaker clustering based on hierarchically structured utterance-oriented dirichlet process mixture model. In *INTERSPEECH*, 2012.