FULLY BAYESIAN INFERENCE OF MULTI-MIXTURE GAUSSIAN MODEL AND ITS EVALUATION USING SPEAKER CLUSTERING

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

[1]Department of Science and Engineering, Waseda University, Tokyo, Japan
[2]Waseda Institute for Advanced Study, Tokyo, Japan
[3]NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan

ABSTRACT

This study aims to verify effective optimization methods for estimating parametric, fully Bayesian models in speech processing. For that purpose, we investigate the impact of the difference in optimization methods for the multi-scale Gaussian mixture model, which is suitable for speaker clustering, on the clustering accuracy. The Markov chain Monte Carlo (MCMC)-based method was compared with the variational Bayesian method in the speaker clustering experiment; with a small amount of data, the MCMC-based method was more effective; with large scale data (more than one million samples), the difference between these methods in terms of the clustering accuracy decreased and the MCMC-based method was computationally efficient.

Index Terms— Speaker clustering, multi-scale Gaussian mixture model, Gibbs sampling, variational Bayesian method

1. INTRODUCTION

Speaker variability in speech data has mainly two levels: 1) inter-utterance variability, which is derived from the difference in speaker characteristics, and 2) intra-utterance variability (i.e., intra-speaker variability), which is derived from the difference in contents of spoken utterances.

Statistical modeling plays an important role in handling data involving such multi-scale properties. In the domain of natural language processing, latent Dirichlet allocation (LDA) [1, 2] is one of the successful approaches for handling multi-scale properties such as word-level and document-level properties. Hierarchical modeling has also been applied to acoustic speaker clustering in the form of an utterance generative model [3, 4]. In this model, we used conventional Gaussian mixture models (GMMs) for representing the intra-utterance variability (i.e., intra-speaker variability) and a mixture of these GMMs for representing the inter-utterance variability (i.e., entire speaker space).

The Markov chain Monte Carlo (MCMC)-based method and the variational Bayesian (VB) method have been applied to the optimization of hierarchical models such as the LDA and utterance generative models. The MCMC-based and VB methods used for LDA were compared previously [2], and it was found that the MCMC-based method is more efficient and takes less computational time to converge parameters. In addition, the VB method is more likely to converge the parameters to the local optima, which yield low-quality solutions. The MCMC-based method, in contrast, can avoid the local convergence problem and improve the quality of the solutions by taking sufficient time for model estimation. The MCMC-based method has been widely used in practice because of its effectiveness, computational efficiency, versatility, and ease of implementation. In this case, discrete data are handled in LDA, whereas continuous data are handled in speech models such as hidden Markov models (HMMs) and GMMs. Hierarchical models are severely affected by local optima, and there is a strong likelihood of this problem being more serious especially in the case of speech modeling that uses continuous Gaussian mixture distributions with a complicated structure.

Therefore, this paper verifies optimization methods suitable for estimating the utterance generative model, which is a hierarchical model for continuous data, in terms of performance and data scalability. For this purpose, we investigate the impact of the aforementioned model optimization methods (i.e., the MCMC-based and VB methods) on the speaker clustering performance by using two speech databases, TIMIT as common speech data and "corpus of spontaneous Japanese (CSJ)" as large scale real speech data.

2. MULTI-SCALE MIXTURES OF GAUSSIAN MODELS

Let $\mathbf{o}_{ut} \in \mathcal{R}^D$ be a D-dimensional observation vector at the t-th frame in the u-th utterance, $\mathbf{O}_u \stackrel{\triangle}{=} \{\mathbf{o}_{ut}\}_{t=1}^{T_u}$ be the u-th utterance that comprises T_u observation vectors, and $\mathbf{O} \stackrel{\triangle}{=} \{\mathbf{O}_u\}_{u=1}^{U}$ be a set of U utterances.

We define the generative model to represent the speaker space by using a mixture of GMMs (MoGMMs) in which D-dimensional GMMs represent speaker characteristics (i.e., intra-speaker variability), and the mixture of these GMMs represents the entire speaker space (i.e., inter-speaker variability). In this model, the number of mixtures in the MoGMMs indicates the number of speakers. To deal with this hierarchical mixture model, two kinds of latent variables are introduced: $\mathbf{Z} = \{z_u\}_{u=1}^{U}$ represents the utterance-level latent variables, each of which identifies the MoGMM component (i.e., speaker distribution) to which the u-th utterance is assigned; $\mathbf{V} = \{v_{ut}\}_{u,t=1}^{U,T_u}$ represents the frame-level latent variables, each of which identifies the intra-speaker GMM component to which the t-th frame in the u-th utterance is assigned. In this case, the utterance-level and frame-level latent variables in the MoGMM-based speech modeling correspond to the document-level and word-level latent variables in LDA, although continuous data are used in speech modeling and discrete data are used in LDA. The conditional probability of all utterances given the latent variables is described as follows:

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

where $\mathbf{\Theta}$ denotes a set of parameters $\left\{\{h_j\}, \{w_{ij}\}, \{\boldsymbol{\mu}_{ij}\}, \{\mathbf{\Sigma}_{ij}\}\right\}$; h_j, the weight for the entire speaker MoGMM component; w_{ij}, $\boldsymbol{\mu}_{ij}$, and $\mathbf{\Sigma}_{ij}$, the weight, mean vector, and covariance matrix for the intra-speaker GMM component, respectively. $\mathbf{\Sigma}_{ij}$ is a diagonal covariance matrix whose (d,d)-th element is represented by $\sigma_{ij,d}$. In

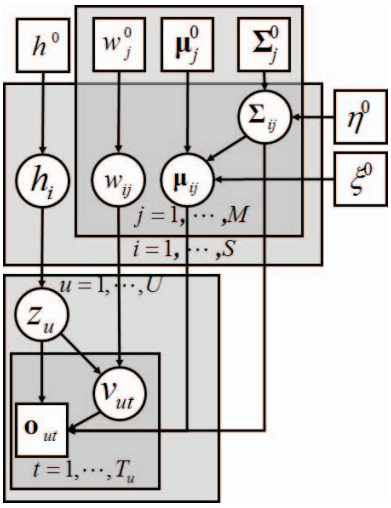

Fig. 1. Graphical model for utterance generative model.

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

$$P(\mathbf{h}) = \mathcal{D}(\mathbf{h}^0), \qquad P(w_{ij}) = \mathcal{D}(\mathbf{w}_j^0),$$
$$p(\mu_{ij,d}, \sigma_{ij,d}) = \prod_{d=1}^{D} \mathcal{NG}(\xi^0, \eta^0, \mu_{j,d}^0, \sigma_{j,d}^0) \quad (2)$$

where $\mathcal{D}\left(\mathbf{h}^0\right)$ denotes the Dirichlet distribution with a hyper parameter $\mathbf{h}^0$ and $\mathcal{NG}\left(\xi^0, \eta^0, \mu_{j,d}^0, \sigma_{j,d}^0\right)$ denotes the Normal-Gamma distribution with hyper parameters ξ^0, η^0, $\mu_{j,d}^0$ and $\sigma_{j,d}^0$. Figure 1 shows the graphical model for the utterance generative model used.

3. MODEL ESTIMATION

When using the multi-scale mixture model described in the previous section, the speaker clustering problem reduces to estimating the utterance-level latent variables. In this case, the following posterior probabilities of the latent variables $\mathbf{V}$ and $\mathbf{Z}$ are estimated.

$$\gamma_{v_{ut}=j|z_u=i} \triangleq p(v_{ut}=j|\mathbf{O}, \boldsymbol{\Theta}, z_u=i) \quad (3)$$
$$\gamma_{z_u=i} \triangleq p(z_u=i|\mathbf{O}, \boldsymbol{\Theta}) \quad (4)$$

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

$$\begin{cases} c_i & = \sum_u \gamma_{z_u=i} \\ n_{ij} & = \sum_{u,t} \gamma_{v_{ut}=j|z_u=i} \cdot \gamma_{z_u=i} \\ \mathbf{m}_{ij} & = \sum_{u,t} \gamma_{v_{ut}=j|z_u=i} \cdot \gamma_{z_u=i} \cdot \mathbf{o}_{ut} \\ r_{ij,d} & = \sum_{u,t} \gamma_{v_{ut}=j|z_u=i} \cdot \gamma_{z_u=i} \cdot (o_{ut,d})^2 \end{cases} \quad (5)$$

where c_i denotes the number of utterances assigned to the i-th component of the entire speaker MoGMM; n_{ij}, the number of frames assigned to the j-th component of the intra-speaker GMM of the i-th component of the MoGMM; and $\mathbf{m}_{ij}$ and r_{ij}, the first and second order sufficient statistics, respectively. The hyper parameters of the posterior distributions for $\boldsymbol{\Theta}$ are computed as follows:

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

In 3.1 and 3.2, we describe the MCMC-based and VB methods, respectively.

Algorithm 1 Collapsed Gibbs sampling-based model estimation.

```
Initialize {z_u, v_ut : u = 1,···,U, t = 1,···,T_u}.
repeat
  for all u such that 1 ≤ u ≤ U do
    for all t such that 1 ≤ t ≤ T_u do
      Sample v_ut from Eq. 7
    end for
  end for
  for all u such that 1 ≤ u ≤ U do
    Sample z_u from Eq. 8
  end for
until some condition is met
```

3.1. MCMC-based method

In the MCMC-based method for estimating the utterance-level latent variables, the samples of latent variables are obtained directly from the posterior distribution of these variables. It means that $\gamma_{v_{ut}=j|z_u=i}$ and $\gamma_{z_u=i}$ described in Eqs. 3 and 4 are zero-or-one values according to the assignment of data. In this study, we apply the collapsed Gibbs sampling, in which the parameters $\boldsymbol{\Theta}$ are marginalized, to the marginalized joint posterior distribution.

In each step of the collapsed Gibbs sampling, the value of one of the latent variables (e.g., z_u) is replaced with a value generated from the distribution of that variable given the values of the remaining latent variables (i.e., $\boldsymbol{Z}_{\backslash u} = \{z_{u'}|u' \neq u\}$). In this case, the latent variables are sampled from the conditional posterior distribution as follows:

[Frame-level latent variables]

$$p(v_{ut}=j'|\mathbf{O}, \mathbf{V}_{\backslash t}, \mathbf{Z}_{\backslash u}, z_u=i) = \frac{\exp\left(g_{ij'}(\tilde{\boldsymbol{\Theta}}_{i,j'}) - g_{ij'}(\tilde{\boldsymbol{\Theta}}_{i,j'\backslash t})\right)}{\exp\left(\sum_j g_{ij}(\tilde{\boldsymbol{\Theta}}_{i,j}) - g_{ij}(\tilde{\boldsymbol{\Theta}}_{i,j\backslash t})\right)} \quad (7)$$

[Utterance-level latent variable]

$$p(z_u=i'|\mathbf{O}, \mathbf{V}, \mathbf{Z}_{\backslash u}) = \frac{\exp\left(\log\frac{\Gamma(\sum_j \tilde{w}_{i'\backslash u,j})}{\Gamma(\sum_j \tilde{w}_{i',j})} + \sum_j \left(g_{ij}(\tilde{\boldsymbol{\Theta}}_{i',j}) - g_{ij}(\tilde{\boldsymbol{\Theta}}_{i'\backslash u,j})\right)\right)}{\exp\left(\sum_i \left(\log\frac{\Gamma(\sum_j \tilde{w}_{i'\backslash u,j})}{\Gamma(\sum_j \tilde{w}_{i',j})} + \sum_j \left(g_{ij}(\tilde{\boldsymbol{\Theta}}_{i,j}) - g_{ij}(\tilde{\boldsymbol{\Theta}}_{i\backslash u,j})\right)\right)\right)} \quad (8)$$

where $g_{ij}(\tilde{\boldsymbol{\Theta}}_{i,j})$ denotes the joint probability described as follows:

$$g_{ij}(\tilde{\boldsymbol{\Theta}}_{i,j}) \triangleq p(\mathbf{O}, \mathbf{V}, v_{ut}=j, \mathbf{Z}, z_u=i) = \log\Gamma(\tilde{w}_{ij}) - \frac{D}{2}\log\tilde{\xi}_{ij} + D\log\Gamma\left(\frac{\tilde{\eta}_{ij}}{2}\right) - \frac{\tilde{\eta}_{ij}}{2}\sum_d \log\tilde{\sigma}_{ij,d} \quad (9)$$

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; and $g_{ij}(\tilde{\boldsymbol{\Theta}}_{i,j\backslash t})$ is computed using $\mathbf{O}_{\backslash t}$, $\mathbf{Z}$, and $\mathbf{V}_{\backslash t}$.

This sampling process is iterated across all latent variables. For the utterance generative model used, the collapsed Gibbs sampling procedure is carried out as Algorithm 1. It should be noted that the sequence of the latent variables sampled is guaranteed to be the sequence of the samples generated from the original joint distribution. We, therefore, can assume that these estimates of the utterance-level latent variables are a result of speaker clustering.

Algorithm 2 Variational Bayesian model estimation.

Initialize $\{z_u, v_{ut} : u = 1, \cdots, U, t = 1, \cdots, T_u\}$.
repeat
  for all i, j such that $1 \leq i \leq S, 1 \leq j \leq M$ **do**
    Compute the expectation values described in Eqs. 16 - 19.
    for all u, t such that $1 \leq u \leq U, 1 \leq t \leq T_u$ **do**
      Compute $q(\mathbf{V}, \mathbf{Z})$ in Eq. 10 followed by computing the expectation values described in Eqs. 13 and 15.
    end for
  end for
  for all i, j such that $1 \leq i \leq S, 1 \leq j \leq M$ **do**
    Compute the hyper parameters of $q(\mathbf{\Theta})$ in Eq. 11 by using the sufficient statistics, as described in Eq. 6.
  end for
until estimation is converged

3.2. Variational Bayesian (VB) method

In the VB method, the utterance-level latent variables are deterministically obtained by estimating the variational posterior distribution, whereas in the MCMC-based method, these variables are stochastically sampled from the conditional posterior distribution. In the VB method, to optimize the variational posterior distribution, we attempt to maximize what is called "free energy," which is the lower bound of the marginalized logarithmic likelihood (i.e., $\log p(\mathbf{O})$).

Under the assumption that each variable in the variational posterior distribution is i.i.d. as $q(\mathbf{V}, \mathbf{Z}, \mathbf{\Theta}) = q(\mathbf{Z})q(\mathbf{V}|\mathbf{Z})q(\mathbf{\Theta})$, the optimal variational posterior distribution (i.e., $q(\mathbf{V}, \mathbf{Z}, \mathbf{\Theta})$ that maximizes the free energy) can be determined as follows:

$$q(\mathbf{V}, \mathbf{Z}) \propto \exp\left(\left\langle \log p(\mathbf{O}, \mathbf{V}, \mathbf{Z}, \mathbf{\Theta})\right\rangle_{q(\mathbf{\Theta})}\right) \quad (10)$$

$$q(\mathbf{\Theta}) \propto \exp\left(\left\langle \log p(\mathbf{O}, \mathbf{V}, \mathbf{Z}, \mathbf{\Theta})\right\rangle_{q(\mathbf{V},\mathbf{Z})}\right) \quad (11)$$

where $\langle f(X)\rangle_{q(X)}$ denotes the expectation given that X is distributed according to q. Optimal $q(\mathbf{V}, \mathbf{Z})$ and $q(\mathbf{\Theta})$ are obtained from Algorithm 2. The posterior probability of a frame-level latent variable is estimated as follows:

$$\begin{aligned}\gamma^*_{v_{ut}=j|z_u=i} &= \exp\Bigg(\langle \log w_{ij}\rangle_{q(w_{ij})} + \frac{1}{2}\sum\nolimits_d \langle \log \sigma_{ij,d}\rangle_{q(\sigma_{ij,d})} \\ &\quad -\frac{D}{2}\log 2\pi - \frac{1}{2}\sum\nolimits_d \langle \sigma^{-1}_{ij,d}(o_{ut,d} - \mu_{ij,d})^2\rangle_{q(\mu_{ij,d}|\sigma_{ij,d})}\Bigg) \quad (12)\end{aligned}$$

In this case, we can determine a frame-level latent variable by normalizing Eq. 12 as follows:

$$\gamma_{v_{ut}=j|z_u=i} = \frac{\gamma^*_{v_{ut}=j|z_u=i}}{\sum_j \gamma^*_{v_{ut}=j|z_u=i}} \quad (13)$$

In the same manner, we can compute an utterance-level latent variable $\gamma_{z_u=i}$ from the posterior probability $\gamma^*_{z_u=i}$ as follows:

$$\gamma^*_{z_u=i} = \langle \log h_i\rangle_{q(h_i)} \prod\nolimits_t \sum\nolimits_j \gamma^*_{v_{ut}=j|z_u=i} \quad (14)$$

$$\gamma_{z_u=i} = \frac{\gamma^*_{z_u=i}}{\sum_i \gamma^*_{z_u=i}} \quad (15)$$

In this case, the expected values of the parameters described in Eqs. 12 and 14 are computed as follows:

$$\langle \log h_i\rangle_{q(h_i)} = \psi(\tilde{h}_i) - \psi(\sum\nolimits_i \tilde{h}_i) \quad (16)$$

$$\langle \log w_{ij}\rangle_{q(w_{ij})} = \psi(\tilde{w}_{ij}) - \psi(\sum\nolimits_j \tilde{w}_{ij}) \quad (17)$$

$$\langle \log \sigma_{ij,d}\rangle_{q(\sigma_{ij,d})} = \psi(\tilde{\eta_{ij}}) - \log \tilde{\sigma}_{ij,d} \quad (18)$$

$$\langle \sigma^{-1}_{ij,d}(o_{ut,d} - \mu_{ij,d})^2\rangle_{q(\mu_{ij,d}|\sigma_{ij,d})} = \tilde{\eta}_{ij}\tilde{\sigma}^{-1}_{ij,d}(o_{ut,d} - \tilde{\mu}_{ij,d})^2 + \tilde{\xi}_{ij} \quad (19)$$

where $\psi(\cdot)$ denotes Digamma function.

Table 1. Number of speakers and utterances in evaluation data. The number of utterances are averaged over speakers in CSJ-4 and 5.

	TIMIT	CSJ-1	CSJ-2	CSJ-3	CSJ-4	CSJ-5
# spkr.	24	10	10	10	10	20
# utt. / spkr.	8	5	10	20	249.1	232.1

4. SPEAKER CLUSTERING EXPERIMENTS

In order to compare the MCMC-based and VB methods, we carried out speaker clustering experiments.

4.1. Experimental condition

4.1.1. Speech data

We performed speaker clustering experiments by using six evaluation sets obtained from the TIMIT and CSJ databases. Table 1 lists the number of speakers and utterances in the evaluation sets used. We used a core test set in TIMIT, which included 192 utterances spoken by 24 speakers. The remaining five evaluation sets were obtained from CSJ as follows: all lectures were divided into utterance units on the basis of silence segments in their transcriptions that were longer than 500 ms; ten speakers were randomly selected and their 5, 10, 20, and all utterances were selected for CSJ-1, CSJ-2, CSJ-3, and CSJ-4, respectively; and 20 speakers were randomly selected and all their utterances were selected for CSJ-5. Each utterance is between 5 and 10 s long. We evaluated five combinations of different speakers in each data set. In this case, CSJ-4 and CSJ-5 include large scale data (about 1M and 3M samples, respectively).

We used 39-dimensional acoustic feature parameters that consisted of 12-dimensional mel-frequency cepstrum coefficients (MFCCs), log energy, their Δ parameters, and their $\Delta\Delta$ parameters. The frame length and frame shift were 25 ms and 10 ms, respectively.

4.1.2. Measurement

We applied the average cluster purity (ACP), average speaker purity (ASP), and geometric mean of those values (K value) to the evaluation criteria in the speaker clustering [5]. The number of iterations in the MCMC-based method was set to 50. We considered the first 40 iterations as the burn-in period, and the K values obtained from this period were rejected. The average of the K values of the remaining ten iterations was measured. Furthermore, we carried out 50 similar experiments using different seeds to generate the random numbers and different initial values for the latent variables and then measured the average of the K values. For the VB method, we also carried out 50 similar experiments using different initial values for the latent variables and then measured the average of the K values.

4.1.3. Evaluation condition

The hyper parameters in Eq. 2 were set as follows: $h^0 = 1$, $\xi^0 = 1$, and $\eta^0 = 1$. w^0_i, $\boldsymbol{\mu}^0_{ij}$, and $\mathbf{\Sigma}^0_{ij}$ in Eq. 2 were set to the weights, mean vectors, and covariance matrices of the universal background model (UBM), respectively. The UBM was trained with the complete test set in TIMIT except for the core test set. The number of speaker clusters was set to the actual number of speakers. The number of mixtures in the intra-speaker GMMs was set to two in both the MCMC-based and VB methods. The initial values of the utterance-level latent variables and those of the frame-level latent variables were determined with the K-means clustering algorithm and assignment of random numbers to the variables, respectively.

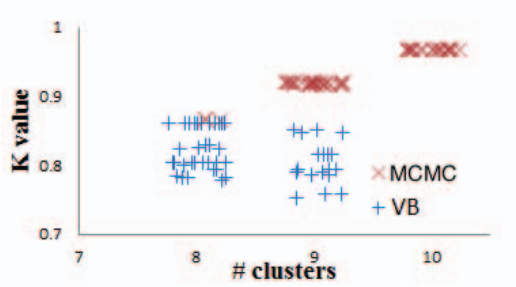

Fig. 2. K values as a function of the number of estimated speaker clusters. To distinguish between data for the same number of clusters, the dots are identified with small random horizontal perturbations.

4.2. Experimental results and discussion

Table 2 lists the ACPs, ASPs, and K values obtained when using the MCMC-based and VB methods for the six evaluation sets. Figure 2 shows a scatter plot representing the relationship between the estimated number of speaker clusters and the corresponding K values for CSJ-1. In this figure, the dots are described with small random horizontal perturbations to distinguish between data for the same estimated number of clusters.

The MCMC-based method outperformed the VB method for most of all evaluation sets. In this case, we can see that the difference in the clustering accuracy between these methods decreased with an increase in the number of utterances for each speaker from the results for CSJ-1, CSJ-2, CSJ-3, and CSJ-4. In addition, the VB method tended to estimate the fewer number of clusters than the true number of speakers, as compared with the MCMC-based method. This tendency became noticeable, especially when the number of utterances was fewer. The VB method is known to easily fall into local optima, whereas the MCMC-based method can avoid such solutions. This hypothesis was supported by Fig. 2; when applying the VB method, the K values were vertically distributed in wide ranges for the same estimated number of clusters, whereas the K values were almost the same in the MCMC-based method. Therefore, the MCMC-based method could achieve more precise speaker clustering than the VB method for small numbers of utterances.

Next, we discuss the reason why the small number of clusters was estimated (in this case, adequate clustering could not be achieved) when the number of utterances was small, focusing on the MCMC-based method. The case where the estimated number of clusters was smaller than the true number of speakers indicates that there were the clusters to which no data were assigned as a result of the model estimation. The results can be interpreted as follows: once a speaker cluster becomes empty (i.e., no data are assigned to this cluster) at the step of sampling the utterance-level latent variables, new utterances are never assigned to this cluster. This indicates that ergodicity in the Gibbs sampler is not satisfied. In this case, it is not guaranteed that the samples are generated from the true posterior distribution. When the number of utterances is large, such a situation will not occur frequently and adequate latent variables would be obtained. In contrast, when the number of utterances is small, the number of speaker clusters to which no utterances are assigned will probably increase, and therefore, adequate clustering would not be achieved.

We finally discuss computational cost. In the experiment performed using CSJ-5 (i.e., 20 speakers and 232.1 utterances per speaker), the MCMC-based method required double times of the VB-based method for one epoch of iterative calculation e.g., the former and latter methods took about 214.87 and 103.01 s, respectively, on an average by using Intel Xeon 3.00 GHz. However, the MCMC-based and VB methods required 10 and 100 iterations until convergence of estimation. As the number of utterances increases, the computational cost will drastically increase because a lot of iterations are needed. Therefore, fast convergence speed of the MCMC-based method has a great advantage in total computational cost. In addition, we can apply some sampling techniques with fast convergence e.g., blocked Gibbs sampling, simulated annealing, and beam sampling.

Table 2. Speaker clustering results.

Evaluation data	Method	#clusters	ACP	ASP	K value
TIMIT	MCMC	23.7	0.783	0.816	0.799
(spkr:24,utt:8)	VB	22.0	0.608	0.790	0.692
CSJ-1	MCMC	9.21	0.808	0.898	0.851
(spkr:10,utt:5)	VB	8.79	0.704	0.860	0.777
CSJ-2	MCMC	9.75	0.852	0.892	0.871
(spkr:10,utt:10)	VB	9.29	0.695	0.846	0.782
CSJ-3	MCMC	9.97	0.866	0.892	0.879
(spkr:10,utt:20)	VB	9.59	0.780	0.870	0.823
CSJ-4	MCMC	10	0.784	0.694	0.738
(spkr:10,utt:249.1)	VB	10	0.773	0.673	0.721
CSJ-5	MCMC	20	0.740	0.627	0.681
(spkr:20,utt:232.1)	VB	18.88	0.693	0.676	0.684

5. CONCLUSION AND FUTURE WORKS

In this study, we investigated the impact of the difference in estimation methods, the MCMC-based and VB methods, of the fully Bayesian multi-scale mixture model on speaker clustering accuracy. Speaker clustering experiments showed that the MCMC-based method outperformed the VB method, especially when only few utterances could be used. We also showed that precise speaker clustering was not always achieved for a data set with a small number of utterances even when using the MCMC-based method, because the parametric Bayesian approach did not always satisfy the ergodicity required in the Gibbs sampler.

In order to solve this problem, it is effective to apply non-parametric Bayesian modeling where the empty clusters can be selected in the Gibbs sampling procedure (e.g., Chinese restaurant process). Actually, in [5], we proposed a non-parametric Bayesian version of an utterance generative model and showed that this model was effective in estimating the number of speakers. In this model, however, the intra-speaker distribution was not represented by an MoGMM but approximately represented by a single Gaussian distribution. Therefore, in future work, we would like to extend the MoGMM used in this study to a non-parametric Bayesian model.

6. REFERENCES

[1] D. M. Blei *et al.*, "Latent Dirichlet allocation.," Journal of Machine Learning Research, vol. 3, pp.993–1022, 2003.

[2] T. L. Griffiths and M. Steyvers. "Finding scientific topics, " Proc. the National Academy of Sciences of the United States of America, vol.101, pp.5228–5235, 2004.

[3] S. Watanabe *et al.*, "Gibbs sampling based multi-scale mixture model for speaker clustering," Proc. ICASSP, pp.4524–4527, May 2011.

[4] F. Valente and C. Wellekens, "Variational Bayesian speaker clustering," Proc. ODYSSEY, The Speaker and Language Recognition Workshop, May 2004.

[5] N. Tawara *et al.*, "Speaker clustering based on utterance-oriented Dirichlet process mixture model," Proc. Interspeech, pp.2905–2908, Aug. 2011.